Artificial Speech for Intensive Care Unit (ICU) Patients and Laryngectomees

by

Marilyn Adeline Mei Lim

B.E. (Hons 1)

A thesis presented for the degree of
Doctor of Philosophy in the Department of Electrical and Computer Engineering

University of Canterbury
Private Bag 4800, Christchurch, New Zealand
31 July 2005

Abstract

A method and prototype device to provide artificial speech for intensive care unit (ICU) patients and laryngectomees is presented. The method assists these patients to produce natural sounding speech by "mouthing the words". A review of the current communication techniques for these patients is presented. The limitations of these techniques suggests that there is a need for a device that produces natural sounding speech (pitch variation and glottal sound source that resembles the actual glottal pulse generated by the vibrating vocal folds) and a device that is user friendly. As vocal folds only vibrate during vowel production, only vowel sounds are considered.

Since pitch variation plays a major role in the naturalness of a person's voice, a number of alternative (automatic) pitch control techniques were explored. A unique pitch control technique utilising the changes in jaw height when a person "mouth the words" is presented.

The electroglottographic (EGG) signal is used as the glottal sound source signal for this research as the properties of the EGG signal offers a number of advantages compared with other glottal sound source measurement techniques. A new glottal source model known as the twin-bar model, based on EGG measurements from normal volunteers, is also introduced. This model changes the shape of the glottal pulse based on a single parameter: pitch. Perceptual testing of the simulated voice using the twin-bar glottal model and two other well-known models on volunteers showed that the twin-bar model produces more natural sounding voice than the other two models.

A new artificial speech system combining the automatic pitch control technique (jaw height) and the glottal sound source (twin-bar model) was constructed. It also includes a number of extra functions that would further improves the speech produced with this system. Existing technology on a laptop (e.g. serial port communication, bluetooth transceivers and USB port) is utilised for the construction of the prototype, with the laptop as the signal processing unit. The prototype was tested on a normal subject.

Acknowledgements

I am very grateful to my supervisor, Associate Professor Philip Bones and my associate supervisor, Dr. Emily Lin (Department of Communications Disorders, University of Canterbury), for their guidance, encouragement, useful and critical comments, and the constant proof-reading throughout the course of my studies. I would also like to thank my two other associate supervisors from Christchurch Hospital, Dr. Geoffrey Shaw (ICU specialist) and Mr. Richard Dove (biomedical engineer from the Medical Physics and Bioengineering Department) for their support, enthusiasm and great ideas.

A special thanks to Dr. Margaret Maclagan (Communications disorders department, University of Canterbury) and Mr. Darren Ayling (speech and language therapist, Christchurch Hospital) for their help and support with the basics of speech and language, at the beginning of my research.

I am indebted to the technical staff, especially Mr. Philipp Hof, Mr. Dudley Berry and Mr. Dermot Sallis at Electrical and Computer Engineering Department who are always ready to lend a hand and for their practical suggestions.

I acknowledge the financial support from the University of Canterbury, the Royal Society of New Zealand, the Royal Society of New Zealand (Canterbury Branch) and the Canterbury Medical Research Foundation.

I am especially grateful to my friends and colleagues for their encouragement and for just being there to listen, at times when my thesis seems to have stalled.

I thank God for guiding me through the ups and downs of this project and for giving me the passion to do a project that I hope will help the less fortunate individuals who cannot speak.

Last, but not least, I would like to thank my wonderful family, especially Christopher, for helping me make this project a reality. I thank them also for their unconditional love and support, for their suggestions and even proof-reading. I could not have done it without them.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Current practice in ICU 2

1.3 Project objectives 2

1.4 Thesis overview 3

1.5 Journal articles and conference presentations 3

2 Speech sounds and characteristics **5**

2.1 Mechanism of human speech production 5

2.2 Speech sounds and articulations 6

2.3 Source-filter model 9

2.4 Speech analysis 11

2.5 Visualising Speech and Spectral Characteristics of Speech 12

2.6 Sound sources and characteristics 13

2.6.1 Voiced sound source 13

2.6.2 Unvoiced sound source 15

2.6.3 Voice types 16

2.6.4 Fundamental frequency range and pitch 16

2.6.5 Prosodic features, jitter and shimmer 17

2.7 Vocal folds movement measurement techniques 18

3 Communication techniques for speech impaired individuals **21**

3.1 Speech impairment in tracheostomised and ventilator dependent patients 22

3.2 Speech impairment in laryngectomees 23

3.3 Standard communication techniques for speech impaired individuals 23

3.3.1 Non-vocal treatment: unaided communication techniques 23

3.3.2 Non-vocal treatment: aided communication techniques 24

3.3.3 Vocal treatment 26

3.3.4 Alternative augmented communication (AAC) devices 32

3.4 Proposal for improved speech production for tracheostomised individuals and laryngectomees 33

4 EGG analysis techniques **35**

4.1 Threshold method 36

4.2 LF-Model method 37

4.3 Computerised Speech Lab (CSL) and Speech Station2 38

4.4 Modified EGG analysis technique 38

4.4.1 Signal segmentation 41

4.4.2 Drift removal 41

4.4.3 Signal differentiation and average frequency calculation 42

4.4.4 Average waveform calculation 46

4.4.5 Parameters extraction 48

5 Glottal Pulse Model and Vocal Tract Model **51**

5.1 The analogy between natural and artificial speech 51

5.2 Existing glottal pulse models 51

5.2.1 Two-pole model 52

5.2.2 Rosenberg model 53

5.2.3 LF-model 54

5.2.4 Physical models 57

5.3 The new glottal pulse model: the twin-bar model 60

5.3.1 Description of the model 61
5.3.2 Analysis 62
5.3.3 Choice of parameters 64
5.3.4 Comparison to earlier models 66
5.4 Vocal tract modelling 66
5.4.1 Acoustic tube model 66
5.4.2 The Kelly-Lochbaum structure 68
5.4.3 Half-sample delay Kelly-Lochbaum structure 70
5.4.4 Z-transform of the lossless tube model 72
5.4.5 Lip radiation 74
5.4.6 Choice of vocal tract model for voice synthesis 74
5.5 Summary 75

6 Alternative pitch control methods for the voiceless **77**

6.1 Pitch control methods considered 78
6.1.1 Thumb pressure sensor 78
6.1.2 Eyebrow movement 78
6.1.3 Laryngeal muscle movement 79
6.1.4 Laryngeal height 79
6.1.5 Random pitch variation 79
6.1.6 Jaw movement 80
6.2 Preliminary study on vocal folds and jaw movement in voiced sounds 80
6.2.1 Method 81
6.2.2 Instrumentation 82
6.2.3 Procedure 84
6.2.4 Data Analysis 84
6.2.5 Statistical Analysis 88
6.2.6 Results 89

6.2.7 Discussion 93
6.2.8 Conclusions 94
6.3 Jaw movement detector for prototype development 95

7 Contribution of glottal wave shape to natural sounding voiced speech 97
7.1 Waveform shape experiment 98
7.1.1 Setup 98
7.1.2 Procedure 100
7.1.3 Waveform analysis 101
7.1.4 Results and discussion 102
7.1.5 Conclusions 111
7.2 Voice synthesis 112
7.3 Perceptual tests 114
7.3.1 Setup 114
7.3.2 Procedure 117
7.3.3 Statistical analysis 118
7.3.4 Results and discussions 119
7.3.5 Conclusion 122

8 Hardware development for MyVoice, the artificial voice device 123
8.1 *MyVoice1*: the first prototype 123
8.1.1 Design layout 123
8.1.2 Prototype testing 126
8.2 *MyVoice2*: the second prototype 127
8.2.1 The design layout 128
8.2.2 Speech enhancement 132
8.3 Use of *MyVoice2* 132
8.3.1 Prototype testing 139

9 Conclusion and suggestions for future research **141**

9.1 Conclusion 141

9.2 Suggestions for future research 142

9.2.1 Experiments and glottal model 142

9.2.2 Hardware development 144

References **157**

Chapter 1

Introduction

Speech is an aspect of humanity that sets us apart from all other creatures of the earth. Our ability to communicate allows us to share our ideas, emotions and needs with one another. In fact, speech comes so naturally to us that it is often taken for granted until something goes wrong with it.

The speech production mechanism involves the interlinking of a variety of complex systems, including the neurological, phonatory, respiratory, and articulatory systems. When one of these complex physiologies is defective, the ability to control speech production becomes impaired. The level of speech impairment may range from something minor like stuttering and slurred speech to the severe case where an individual becomes incapable of producing any speech or even voice. This thesis looks at the effect of airway obstruction in the intensive care unit (ICU) and laryngectomised patients on speech production and a new method of restoring speech for these patients.

1.1 Motivation

A number of patients who are being treated in ICU due to serious medical illnesses or injury have tracheostomy tubes introduced into them (the insertion of a breathing tube into the trachea) and have to rely on a mechanical ventilator (a machine that pumps air into the lungs) to assist them with their breathing. One of the major problems with these mechanically ventilated patients in ICU is their inability to speak because the airflow necessary to set vocal folds into vibration during phonation or to formulate sounds with speech articulation is directed through the cuffed tracheostomy tube bypassing the modulation of the larynx and the resonance and sound shaping of the nasal and oral tract. The inability to speak may result in anxiety, agony, fear, panic, feeling of helplessness and frustration that can be detrimental to the patient's emotional and physical condition. There are roughly 100,000 people who are treated in ICUs in New Zealand and Australia every year [CBC+02]. These include 1325 patients from Christchurch Hospital. Of all the patients in ICU, approximately 80 - 85% are tracheostomised and dependent on a ventilator to assist them in their breathing.

Another group of patients incapable of producing natural speech are laryngectomised patients, whose larynx (voice box) have been surgically removed due to malignant lesions or serious accident in the neck region. In such cases, air enters and exits the lungs through a hole (stoma) in the neck instead of through the nose or mouth. Hence laryngectomised patients are also unable to speak normally. A study in America estimates that there are over 30,000 laryngectomised individuals living in the United States alone, with approximately 12,000 new cases diagnosed each year. In New Zealand, this number is yet unknown.

Over the years, a number of vocal and non-vocal communication techniques have been developed to help these patients to communicate. However, many of these techniques require a lot of time and effort from the patients in order to convey a simple message, causing them to experience frustration and feelings of inadequacy. Others may be unsuitable due to the patients' physical condition.

In light of the difficulties encountered with the current methods of communication, a new communication technique has been researched where a voice prosthesis is used to allow these patients to produce (near) normal speech even in the absence of the voice box. Such a device could greatly enhance the quality of life for the patients and make it easier for staff to better assess their need. A literature review on different types of voice prothesis is provided, including some evaluation of their advantages and disadvantages.

1.2 Current practice in ICU

Due to the physical state of ICU patients, most communication between patients and their caregivers are usually limited to non-vocal techniques such as alphabet/picture boards, lip-reading, facial expression and hand-grasping/eye blinking [Mas93]. These techniques require some prompting from the caregiver. Some patients who are strong enough may be able to use pencil and paper. Electrolarynx is another option but this requires some learning and good hand coordination, which may be less suitable for these patients. In addition, the quality of the voice produced is quite poor, almost robot-like due to the single-pitch signal produced by the device. Details of the common methods of communication for ICU and laryngectomised patients with their caregivers are given in Section 3.3.

1.3 Project objectives

The aim of this project is to create a device that will allow patients in Intensive Care Unit and laryngectomees who are unable to speak by natural means due to airway obstruction to produce natural sounding speech by "mouthing the words". It is essentially an artificial larynx that incorporates natural control of speech signals from an intact part of the body to achieve the desired qualities of a normal larynx. Emphasis is also placed on making the device user friendly.

1.4 Thesis overview

The remainder of the thesis is outlined below. Original contributions are identified.

Chapter 2 outlines the background information on the physiology of speech, the linguistic aspects as well as the characteristics of speech relevant to the research work reported in this thesis. The source-filter theory of speech production is reviewed.

The various types of communication techniques for speech impaired individuals, including their advantages and disadvantages, are reviewed in Chapter 3.

The current techniques of analysing vocal fold movement are reported in Chapter 4. A new technique specifically designed for the purpose of this project is also presented here.

A review of the various types of glottal pulse models is presented in Chapter 5. The twin-bar model, a new glottal pulse model used as the sound source for this thesis, is described. Vocal tract models used for voice synthesis are also discussed in this chapter.

Chapter 6 investigates alternative methods for controlling the pitch of an artificial speech device, with emphasis on automatic pitch control. A study to investigate the relationship between vocal folds movement and jaw height in voiced sounds, and the design of a new automatic pitch controller for artificial larynx are reported.

Chapter 7 describes the contribution of glottal waveform shape to natural sounding voiced sounds. A study was carried out to find the parameters that determine the waveform shape of the voice source. Equations for the parameters are used in the new glottal pulse model (twin-bar model) to generate the desired glottal pulse. Synthesised voice is generated using the twin-bar model and two other well-known glottal pulse models. Perceptual tests were carried out to test the intelligibility and quality of the synthesised voice using these three glottal models.

The hardware development and prototype for *MyVoice*, the new artificial voice device, are presented in Chapter 8. An initial prototype (*MyVoice1*) was constructed as a proof of concept. *MyVoice2* is a more sophisticated prototype that incorporates the new ideas obtained from a series of studies carried out in this research. Preliminary tests were carried out to assess the quality of speech produced with *MyVoice1* and *MyVoice2*. Finally, conclusions and suggestions for future research are presented in Chapter 9.

1.5 Journal articles and conference presentations

Listed below are papers and conference presentations on the research that I have authored and co-authored but which have not yet been presented for examination for any degree:

Lim M. A., Bones P. J., Lin E., "Twin-Bar Model: An Alternative Glottal Voice Source Model for Voice Simulation". (in preparation for submission to the Journal of Phonetics)

Lin E., Lim M., Bones P., Ormond T., Hornibrook, J., "Comparing Two Demarcation Methods for Electroglottographic Measures of Normal and Pathological Voices." (in preparation for submission to the Journal of the Acoustical Society of America)

Lin E., Ormond T., Hornibrook J., Lim M., "The Effect of Pitch and Loudness on the Electroglottographic Measures in Voice Patients." (in preparation for submission to the Journal of Voice)

Lin E., Bones P., Lim M., "Variations of Glottal Vibratory Patterns in Normal Modal Voice." (in preparation for submission to the Journal of Voice)

Lin E., Ormond T., Ayling D., Lim M., Hornibrook, J., "Effect of Manner of Voice Initiation on Pathological Voice." Journal of Speech, Language, and Hearing Research. (in submission, 8 June, 2005)

Lim M. A., Lin E., Bones P. J., "Vowel Effect on Glottal Parameters and Magnitude of Jaw Opening", Journal of Voice, 2005. (in press)

Conference presentations (national and international meetings)
Lin E., Ormond T., Ayling D., Lim M., Hornibrook J., "Effect of Voice Initiation on Phonatory Quality in Pathological Voice." Poster presentation at the American Speech-Language and Hearing Association's 2004 Annual Convention, Philadelphia, Pennsylvania, USA, November 18-21, 2004.

Lim M. A., Lin E., Bones P. J., "Vowel Effect on Glottal Parameters and Magnitude of Jaw Opening", Presented at the Voice Foundation's 32nd Annual Symposium: Care of the Professional Voice, Philadelphia, Pennsylvania, 4-8 June 2003.

Marilyn Lim, Philip J Bones, Emily Lin, Geoffrey Shaw, Richard Dove, "Artificial Speech for Tracheostomized and Ventilated Patients in Intensive Care Unit (ICU)", Presented at NZPEM 2001 Conference, Australasian College of Physical Scientists and Engineers in Medicine, Christchurch, November, pp. 19.

Chapter 2

Speech sounds and characteristics

This chapter describes the speech sounds produced by humans, particularly by English language speakers. The unique human speech production mechanism, including the characteristics of speech that is produced and the linguistic aspects of speech, are explored.

2.1 Mechanism of human speech production

Speech is comprised of voiced and unvoiced sounds. Voiced sounds such as vowels are produced with vibrating vocal folds while unvoiced sounds such as most consonants are produced without vibrating vocal folds. A more detailed description of voiced and unvoiced sounds can be found in section 2.6. As the focus of this project is on natural sounding artificial voice device (e.g. artificial sound source that mimics the vibration of the vocal folds), the speech production discussed here concentrates on voiced sounds alone.

The onset of speech production begins in the brain where thoughts, emotions and needs are transformed into neural information using an appropriate set of words and grammatical rules learned from an early age. This information is then coded into a sequence of neuromuscular signals (electrical pulses) that is used to control various speech production organs of the body including the vocal folds, vocal tract, articulators (e.g. tongue, lips, jaw) and the respiratory system.

When a person wishes to speak, the electrical pulses from the brain travel through the nerves into the inspiratory muscles (muscles that control inhalation) and signal them to lift the rib cage up and out as well as lowering the diaphragm. This has the effect lowering the pressure inside the lungs causing air to flow into the lungs. Once the lungs are filled, the expiratory muscles contract allowing the rib cage to lower and diaphragm to return to its relaxed elevated state. This increases the pressure inside the lungs, forcing a stream of air to flow from the lungs into the trachea. Although trachea sizes vary from individual to individual, the average trachea size for adults is approximately 110 to 120 mm in

length and 25 mm in diameter [RHI05, PGU98, DM04].

Right at the top of the trachea is the larynx or voice box. Nested inside the larynx are the laryngeal muscles, cartilages, vocal folds, and an opening between the vocal folds known as the glottis. If the vocal folds are closed, the stream of air from the trachea is not able to flow through. The air below the vocal folds is therefore trapped and its pressure starts to build up. The pressure increases until it is high enough to force the vocal folds apart. As the air rushes through the glottis, it creates a phenomenon called the Bernoulli effect where the pressure below the vocal folds drops below the pressure at the top of the folds allowing the tension from the laryngeal muscles to close the glottis. With the glottis closed, once again the pressure below the vocal folds starts to build up. The process repeats, forming a stream of quasi-periodic skewed-sinusoidal pulses with a "buzz-like" sound. This "buzz-like" sound is called the glottal pulse. The fundamental frequency (*F0*) of the glottal pulse is perceived as pitch and it varies depending on the air pressure generated by the lungs and the diameter of the trachea, as well as the tension, length and mass of the vocal folds. (Note: In literature related to speech, 'pitch' means the fundamental frequency of the vocal folds vibration [SS78]. In the hearing field, 'pitch' means the perception or physiological sensation of sound related to the frequency of that sound. In this thesis the terms 'pitch' and 'fundamental frequency' are used synonymously). The diameter of the trachea and mass of the vocal folds are fixed for each individual, while the air pressure from the lungs and the tension and length of the vocal folds vary as the individual speaks.

Beyond the larynx is the pharynx, a flexible structure located above the trachea and behind the nose and mouth. The glottis, pharynx, velum (soft palate), tongue, jaw, teeth, and lips and nasal cavity form the pharyngeal-oral tract, commonly known as the 'vocal tract' [SS78, Pic98]. The average distance from the glottis through to the lips is typically 175mm [Pul05, III94, JHP00]. The length is a parameter used for vocal tract modelling for adult males. The vocal tract acts as resonator for the glottal pulse by amplifying certain harmonics of the glottal pulse and attenuating others. By varying the shape of the vocal tract, and hence the resonators, a wide range of speech sounds can be produced.

2.2 Speech sounds and articulations

Linguists have found that every spoken language can be decomposed into the most basic form of speech sounds known as phonemes [Pic98]. These building blocks of speech sounds in English are related to the vowels and consonants. When one or more phonemes are combined together, they form syllables. A combination of syllables form words and words form sentences. There are 44 phonemes in New Zealand English [Gor96]. Out of the 44 phonemes, 11 are vowels. Figure 2.1 shows the phonemes for the general New Zealand English using the International Phonetic Alphabet [Gor96].

Articulators are responsible for shaping or modulating the sound from the speech source to produce

The symbols used here are those of the International Phonetic Alphabet (IPA). Some are familiar but some need to be learned.

THE PHONEMES OF NEW ZEALAND ENGLISH USING THE INTERNATIONAL PHONETIC ALPHABET

/i/	feed	/fid/	/p/	pin	/pɪn/
/ɪ/	fit	/fɪt/	/b/	bin	/bɪn/
/e/	fed	/fed/	/t/	tin	/tɪn/
/ɔ/	ford	/fɔd/	/d/	din	/dɪn/
/ʊ/	foot	/fʊt/	/g/	got	/gɒt/
/u/	food	/fud/	/k/	cot	/kɒt/
/æ/	fat	/fæt/	/f/	fat	/fæt/
/ɒ/	fog	/fɒg/	/v/	vat	/væt/
/ʌ/	fun	/fʌn/	/ʃ/	ship	/ʃɪp/
/ɜ/	fern	/fɜn/	/ʒ/	beige	/beiʒ/
/a/	farm	/fam/	/h/	hat	/hæt/
			/l/	lit	/lɪt/
/ei/	fate	/feit/	/r/	rat	/ræt/
/ou/	foe	/fou/	/m/	mat	/mæt/
/ai/	fine	/fain/	/n/	not	/nɒt/
/au/	frown	/fraun/	/w/	win	/wɪn/
/ɔi/	foil	/fɔil/	/j/	you	/ju/
			/s/	sue	/su/
/iə/	pier	/piə/	/z/	zoo	/zu/
/eə/	pear	/peə/	/ð/	this	/ðɪs/
/uə/	cure	/kjuə/	/θ/	thick	/θɪk/
			/ŋ/	sing	/sɪŋ/
/ə/	banana	/bənanə/	/tʃ/	church	/tʃɜtʃ/
			/dʒ/	judge	/dʒʌdʒ/

Figure 2.1 Phoneme chart for New Zealand English (reproduced from [Gor96]).

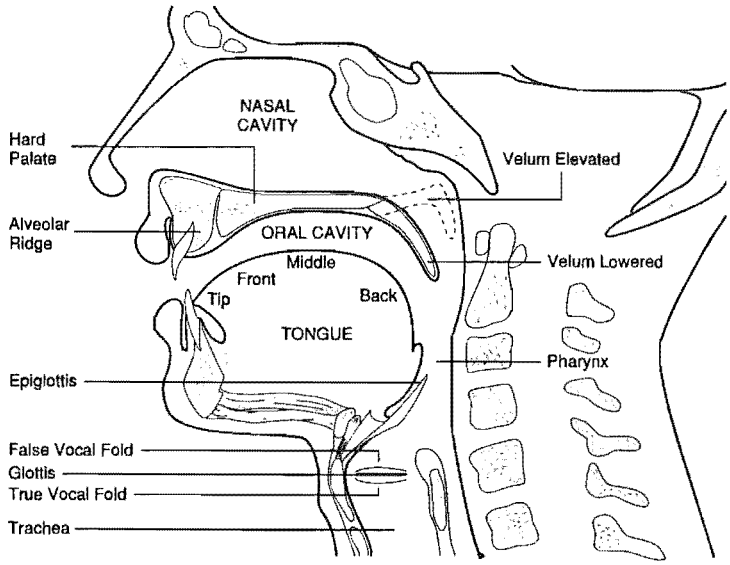

Figure 2.2 The articulators [RE96].

the different speech sounds (see Figure 2.2). Articulators consist mainly of the tongue, lips, hard palate, velum and jaw.

There are two types of vowels: short vowels (/I/, /e/, /ʊ/, /æ/, /ɐ/ and /ʌ/) and long vowels (/i/, /ɔ/, /u/, /ɜ/ and /a/) (refer to Figure 2.1 for examples of the different vowels). Vowels are formed mainly by changing the contour of the tongue. The vocal tract stays relatively open during vowel production. Vowels are classified according to the position of the highest point of the tongue forming the vowel (see Table 2.1) and the lips' shape (e.g. lip-rounding or otherwise). Long vowels are the same as short vowels except that the duration of utterance is sustained for a longer period. Diphthongs (/ei/, /oʊ/, /ai/, /aʊ/, /ɔi/, /iə/, /eə/ and /ʊə/) are created when the articulators move from the position of one vowel to the position of another vowel [Fry79].

Consonants are classified according to the place of articulation and manner of articulation (refer to Table 2.2). The following are the nine basic places of articulation used in the New Zealand English language (refer to Figure 2.1 for the pronunciation of the consonant symbols listed):

- Bilabial (by bringing the lips together, e.g. /b/, /m/, /p/ and /w/)
- Alveolar (the placement of the tip of the tongue on the alveolar ridge, e.g. /t/, /d/, /s/, /z/ and /l/)

Table 2.1 Vowel phonemes for New Zealand English [Mac82]

	Tongue position		
	Front	Central	Back
Close	/i/ peat		/u/ boot
	/ɪ/ pit		/ʊ/ put
	/e/ pet	/ ɜ/ pert	
	/æ/ cat	/ʌ/ but	/ɔ/ port
Open	/a/ part		/ɐ/ pot

- Velar (constriction between back of tongue and soft palate to produce /g/, /k/ and /ŋ/ sounds)
- Labiodental (constriction slit between the lower lip and incisors for the sound /f/ and /v/)
- Interdental (/θ/, /ð/)
- Palato-alveolar (constriction between blade of tongue and hard palate, e.g. /tʃ/, /dz/, /ʃ/, /z/)
- Glottal (used to produce /h/, the only sound in English made by constriction of the vocal folds)
- Post-alveolar (/r/)
- Palatal (/j/)

In terms of the manner of articulation, the phonemes can be divided into 7 categories [BB98]: plosive (e.g. /p/, /b/, /t/, /d/, /k/ and /g/), fricative - made with a narrow constriction such that air creates a noisy sound as it rushes through the narrowed passage (e.g. /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, and /h/), affricate (/tʃ/ and /dz/), nasal (/m/, /n/ and /ŋ/), lateral (/l/), frictionless continuant (/r/) and semi-vowel/glide (/w/ and /j/).

2.3 Source-filter model

The production of speech can be modelled by the source-filter theory of speech production, first developed by Fant [Fan60]. Figure 2.3 shows a simplified model for the production of voiced speech using this model. The source-filter model assumes that the source signal (vocal folds vibration), *g(t)*, is relatively independent of the vocal tract [Lin85]. The vocal tract, *h(t)*, can be modelled as a series

Table 2.2 Consonants classification based on place of articulation and manner of articulation.

	Place of articulation									
Manner of Articulation	Plosive		Fricative		Affricate		Nasal	Lateral	Liquid	Glide
	v	uv	v	uv	v	uv	v	v	v	v
Bilabial	/b/	/p/					/m/			/w/
Alveolar	/d/	/t/	/z/	/s/			/n/	/l/		
Velar	/g/	/k/					/ŋ/			
Labiodental			/v/	/f/						
Interdental			/ð/	/θ/						
Palato-alveolar				/ʃ/	/ʤ/	/ʧ/				
Glottal				/h/						
Post-alveolar			/z/						/r/	
Palatal			/ʒ/		/ʤ/					/j/

Note: v = voiced, uv = unvoiced.

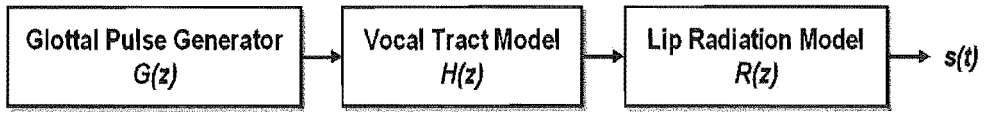

Figure 2.3 Simplified model for voiced speech synthesis using the source-filter model.

of band-pass filters or resonant cavities. The source signal is linearly filtered through the vocal tract where harmonics of source signal that lie near the natural resonances of the vocal tract are reinforced while others are attenuated. Voice, $s(t)$, is produced when the filtered signal passes through the lips, $r(t)$, and is radiated into the air. By altering the shape of the vocal tract, and hence the resonances of the vocal tract, different sounds can be produced.

The simplified source-filter model equation can be written as:

$$s(t) = g(t) \otimes h(t) \otimes r(t) \tag{2.1}$$

where $\otimes$ is the convolution operator. Refer to equation 2.4 for the definition of convolution.

2.4 Speech analysis

All signals, including speech signals, can be represented by a unique combination of sinusoids, each with a different frequency, phase and amplitude. This representation of the speech signal is called the speech spectrum. The spectrum of a speech signal, $S(f)$, can be obtained using the forward Fourier Transform (FT) (equation 2.2). The inverse Fourier Transform (equation 2.3) converts the speech spectrum back to the time domain signal, $s(t)$.

$$S(f) = \int_{-\infty}^{+\infty} s(t) e^{-j2\pi ft} dt \tag{2.2}$$

$$s(t) = \int_{-\infty}^{+\infty} S(f) e^{j2\pi ft} df \tag{2.3}$$

An important operation in signal processing is convolution. The convolution of two continuous functions of time, $g(t)$ and $h(t)$, is given by

$$v(t) = g(t) \otimes h(t) = \int_{-\infty}^{+\infty} g(t')h(t-t')dt' \tag{2.4}$$

If $g(t)$ is the glottal source and $h(t)$ is the vocal tract impulse response, Eqn. 2.4 represents the signal arriving at the lips. The convolution of two signals is equivalent to multiplication of the spectra of the two signals in the Fourier domain (the convolution theorem):

$$\mathscr{F}(g(t) \otimes h(t)) = G(f)H(f) \tag{2.5}$$

where $\mathscr{F}$ is the Fourier Transform operator. For sampled signals, the continuous integral in the FT is replaced by summation. This is called the Discrete Fourier Transform (DFT). Equations 2.6 and 2.7 show the DFT and inverse DFT respectively. N is the number of samples in the signal.

$$G[k] = \frac{1}{N} \sum_{n=0}^{N-1} g[n] e^{-j2\pi nk/N} \tag{2.6}$$

$$g[n] = \sum_{k=0}^{N-1} G[k] e^{j2\pi nk/N} \tag{2.7}$$

The Fast Fourier Transform (FFT) is a DFT algorithm which reduces the number of operations re-

quired to calculate the *DFT* of N points from N^2 to $2N \log_2 N$. MATLAB (MathWorks, Inc.), the software package that is extensively used throughout this thesis, uses the *FFT* for its *DFT* computation.

A more useful representation of the source-filter model is the Z-transform [OS75]. The Z-transform of a digitised signal, $g[n]$ is defined by:

$$G(z) = \sum_{n=0}^{+\infty} g[n] z^{-n} \tag{2.8}$$

where $z = re^{j\omega}$ and $-\pi \leq \omega \leq \pi$. *DFT* of $g[n]$ may be obtained from Z-transform by substituting $z = re^{j\frac{2\pi k}{N}}$ where $r = 1$. *DFT* is therefore a special case of Z-transform evaluated around the unit circle.

Using the Z-transform, the source-filter equation (Eqn. 2.1) can now be written in the form of Eqn. 2.9, where *S(z)* is the Z-transform of *s(t)*, *G(z)* is the Z-transform of *g(t)* and so on.

$$S(z) = G(z)H(z)R(z) \tag{2.9}$$

2.5 Visualising Speech and Spectral Characteristics of Speech

Speech is an abstract entity: it can be perceived but not seen. One method of visualising speech is to measure the acoustic signal with a microphone. A microphone detects the changes in sound pressure level from speech and transforms it into electrical signal that can be plotted onto a graph. Figure 2.4 shows the sound pressure level versus time waveform of the phrase "Say t/a/ again" by a New Zealand speaker. During the vowel segments, the vocal tract is relatively open and there is only a small amount of coupling into the nasal cavity. As a result, vowels contain more speech energy than unvoiced sounds. This effect can be observed in the significantly higher amplitude in the vowel portions of the time domain signal in Figure 2.4.

Figure 2.5 shows the spectra at the various stages of the speech production process. Formants or resonances of the vocal tract are distinct features in the speech spectrum. The positions of the formants, particularly the first three formants, are important in determining the phonetic identity of a speech sound [Fan73]. Singing sounds require higher formants. It is worth noting that unlike the radiated speech spectrum, the glottal source spectrum and the vocal tract spectrum are not generated from recorded signals but through calculations made from models (e.g. inverse filtering of acoustic signal or airflow signal [JABM87, LMJ88, Rot73] and, vocal tract model from vocal tract area function [ST96]).

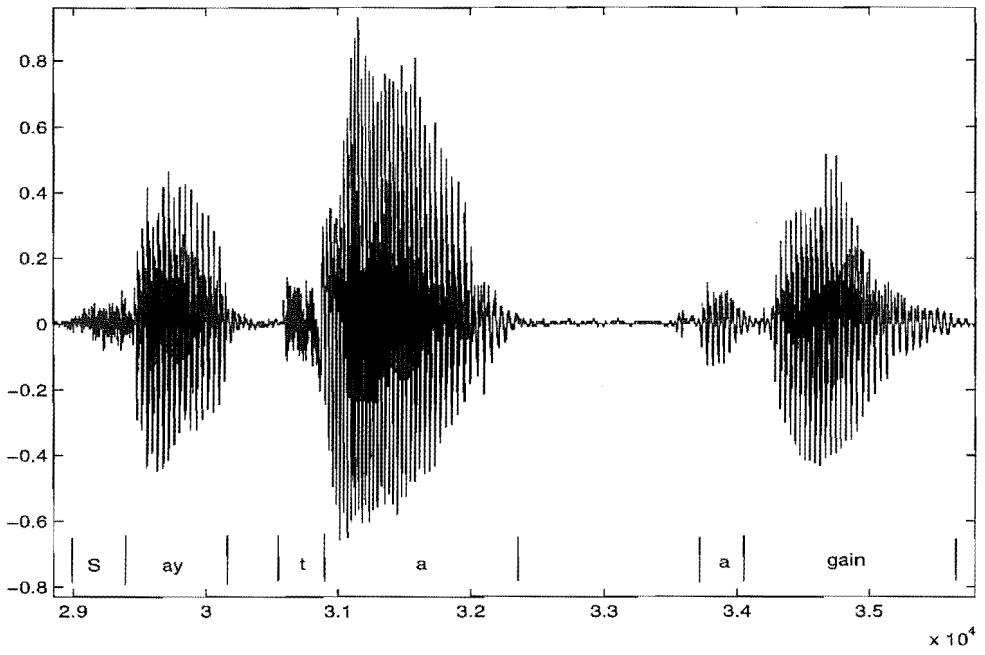

Figure 2.4 Acoustic waveform for the phrase "Say t/a/ again"

The spectrogram is another useful method of visualising speech. As illustrated in Figure 2.6, the spectrogram of a signal is presented as the magnitude of the time-dependent Fourier transform versus time. It allows the observation of the frequency content and intensity of a speech signal at a particular time of the speech utterance. As with speech spectrum, other features that can be observed in the spectrogram include the fundamental frequency (*F0*), the harmonics of the signal and also the formants.

2.6 Sound sources and characteristics

There are two main categories of sound sources in speech: voiced and unvoiced. Different vibrating patterns of the vocal folds produce different types of sound source which, in turn, influence the quality of voice produced.

2.6.1 Voiced sound source

The voiced sound source (glottal pulse) has already been described in section 2.1: the glottal pulse is generated by the periodic modulation of the expired air stream from the repeated opening and closing of the glottis. Vowels, nasal consonants (/m/,/n/,/ŋ/), liquid consonants (/r/) and glide consonants (/w/,/j/) are some of the examples of sounds produced with voiced sound source.

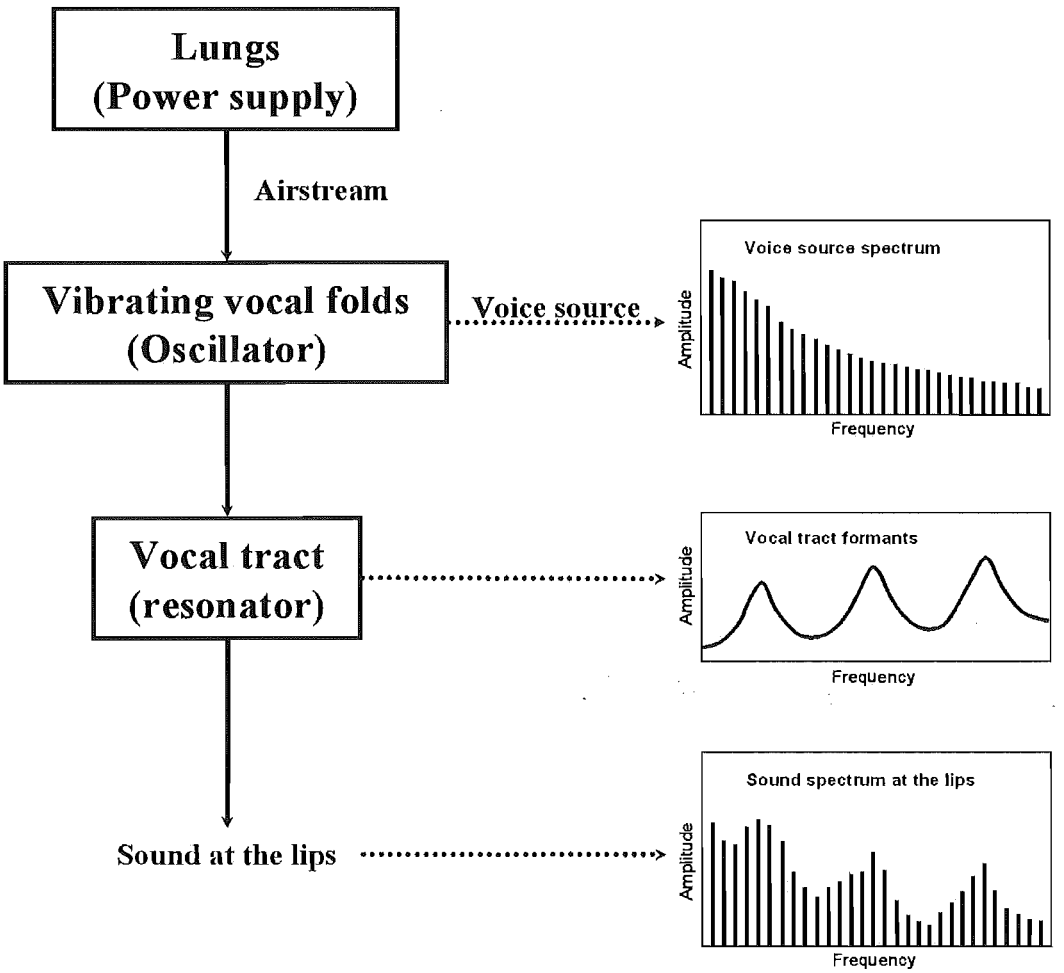

Figure 2.5 The spectra at various stages of the speech production process.

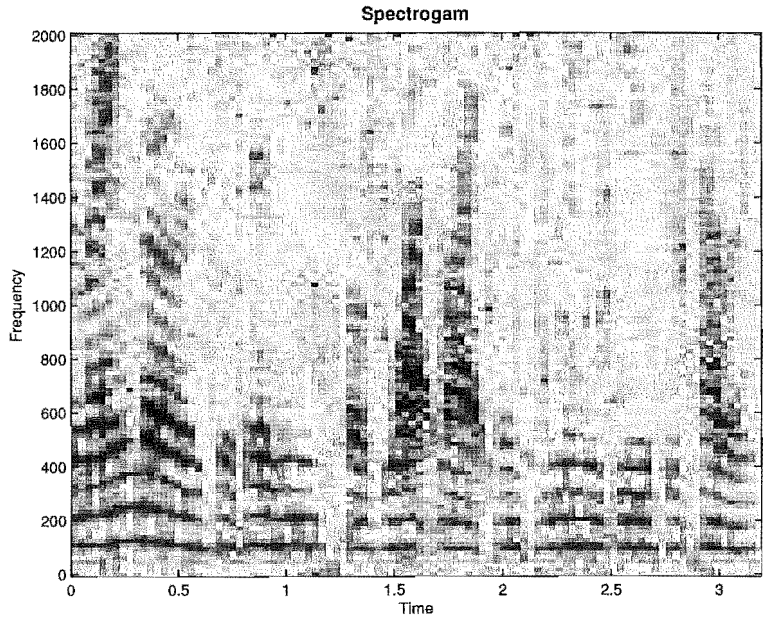

Figure 2.6 The spectrogram.

2.6.2 Unvoiced sound source

For the unvoiced sound source, the vocal folds do not vibrate and the airstream from the lungs flows through the vocal tract in a random or aperiodic manner [Pic98]. Unvoiced sound sources can be further subdivided into turbulent and transient types. A turbulent sound source is generated when airstream is forced through a narrow constriction within the vocal tract, for example when air is expelled between the tongue and the roof of the mouth. It has a 'hiss-like' sound similar to the 'white noise' generated by an un-tuned television. Examples of turbulent sounds are fricatives like (/f/,/ð/,/s/) and (/ʃ/).

A transient sound source is generated when the constriction along the vocal tract is totally closed, stopping the airflow and at the same time allowing the pressure behind the constriction to build up [Pic98]. Sudden releases of the complete constriction produce transient 'clicks' and 'pops'. Speech sounds like (/p/,/t/) and (/k/) are some examples of sounds produced by transient sound sources.

A simple test for determining if a sound is generated by a voiced or unvoiced sound source is to place the thumb and index finger on either side of the thyroid cartilage and utter the sound. If vibration is felt while the sound is uttered, then it is a voiced sound. Likewise, a sound is unvoiced if there is no vibration when the sound is uttered.

The artificial speech device proposed in this thesis focuses on the voiced part of speech. Suggestions for unvoiced speech production is discussed in the future research section in Chapter 9.

2.6.3 Voice types

Voice has been classified in a variety of ways based on perceptually distinguishable quality, such as registers. Vocal registers refer to regions of voice productions along the pitch continuum that are perceived to be associated with different manners of laryngeal adjustment for phonation. There are 3 major voice types for (normal) human speech: modal, falsetto and vocal fry registers [Hol74]. Others include rough, harsh, breathy, strident, nasal and whisper.

Modal register, also known as chest register, refers to the manner of laryngeal adjustment and vocal fold vibration used in the production of normal speech. In this register, the vocal folds are thick and short. Each vibrating cycle has a distinct opened and closed phase. In falsetto register, the vocal folds are thin and long, vibrating at the upper extreme of the human vocal frequency range and the glottis is not fully close. Vocal fry register is at the lowest end of the human vocal frequency range [CC96], occurring approximately 1 octave below the average frequency of the male modal register (20-70Hz) with a mean of around 50Hz [BCNG98]. Whitehead *et al* [WMW84] and Moore and von Leden [MvL58] described that during vocal fry vocal folds may have single or multiple opened phases during one vibratory cycle. It has been reported that most people can generate vocal fry but not many routinely and extensively utilise this type of phonation [Hol68]. Whisper mode occurs when the vocal folds do not approximate during phonation (e.g. the glottis is opened).

As different voice registers use different manner of utterance, only modal voice, the most common of all voice types is addressed in this thesis.

2.6.4 Fundamental frequency range and pitch

The habitual pitch for male voice is between 100-150Hz, while for female it is between 180-250Hz [CC96, TE93], both with a standard deviation of utterance at approximately 2-4 semitones from the habitual pitch [BO00, TE93, JL86]. The definition of semitone, n, is shown in Eqn. 2.10. An octave is defined as 12 semitones above or below the pitch of interest (in this case, the habitual pitch). For example, if the habitual pitch of a person is 100Hz, an octave above the habitual pitch will be 200Hz and an octave below the habitual pitch will be 50Hz. The pitch range for an individual may be as high as 3 octaves [CC96, HJ73]. However, untrained individuals are rarely able to sing beyond 2 octaves easily. Figure 2.7 shows the effect of pitch change on the speech spectrum at the lips. When the pitch is high (Figure 2.7, left) such as a female voice, the formants of the radiated spectrum (hence the sound identity) is more difficult to determine. The reason for this is because the number of harmonics that represent the signal is lower compared with the lower *F0* case (Figure 2.7, right).

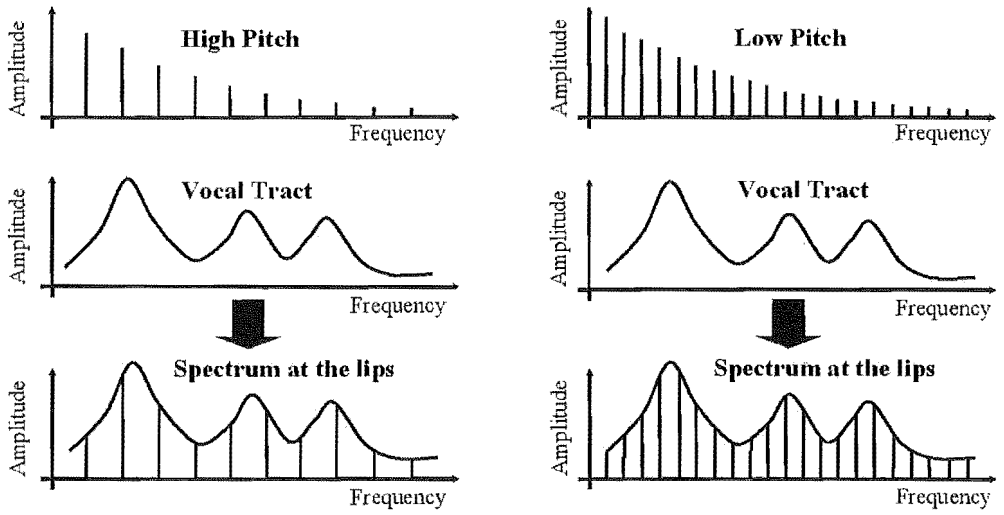

Figure 2.7 Speech spectrum at the lips for high pitch (left) and low pitch (right).

$$n = 39.86 \log_{10} \frac{F1}{F0}, \tag{2.10}$$

where n = number of semitones, $F0$ = habitual pitch and $F1$ = pitch of interest.

2.6.5 Prosodic features, jitter and shimmer

Prosody refers to the stress patterns of an utterance or the long-term properties of speech. The prosodic features of speech consist of *F0*, duration and amplitude/intensity [CP86, Pic98]. In general, the *F0* and amplitude of an unstressed utterance or statement show a downward trend towards the end of the statement [Pic98, Mae76, Lie67]. This could be due to the drop in subglottal pressure towards the end of the statement.

In addition to stress pattern, normal voice shows short-term cycle-to-cycle variation in the fundamental frequency, known as jitter [Tit94]. Shimmer, another inherent characteristic of speech is the short-term variation in the vocal intensity [RE96]. Voice intensity is a measure of the magnitude of sound pressure level and can be written as:

$$\textit{Voice Intensity (dB)} = 20 \log_{10} \frac{P}{2 \times 10^{-5}} \tag{2.11}$$

where numerator, P, is the measured pressure in N/m^2 and the denominator is the reference pressure of the human hearing threshold level [LB88].

Commercial software such as CSL [sl00] can be used to measure the *F0*, duration, loudness, jitter, shimmer and intensity of the recorded EGG and acoustic signals. All these factors affect the naturalness and intelligibility of speech.

2.7 Vocal folds movement measurement techniques

There are many techniques for measuring vocal folds movement. The following briefly describes some of the more commonly used techniques.

The laryngoscope is a device with a camera attached to one end of a flexible fibre-optic cable [Gra97]. Laryngoscopy is performed by inserting the tube into the pharynx via the nose. As this procedure can be quite uncomfortable for the person who is undergoing the procedure, local anaesthesia is required. It is quite commonly used in speech clinics to examine the laryngeal behaviour of a patient.

Ultra high-speed photography as its name stated uses a very high frame rate camera to study the vocal folds movement more precisely [Moo68, Hir88]. This device is expensive and it is invasive as it requires the insertion of a tele-endoscope into the pharynx.

The photoglottograph (PGG) is another type of instrument used for measuring vocal folds movement [Zem88], particularly glottal area functions [Har75]. Measurement is carried out by placing a light source below the vocal folds (just below the cricoid cartilage) and a photo-detector above the vocal folds (in the pharynx). PGG measures the amount of light that passes through the glottis as the vocal folds vibrate. Like the ultra high-speed photography method, the main drawback of PGG is the invasive nature of this technique.

The electroglottograph (EGG) is a non-invasive device that measures the vocal folds contact [Chi84]. EGG provides the vocal folds contact information by measuring the electrical impedance between 2 electrodes placed around the neck. Figure 2.8 shows an example of the EGG device. Since EGG device is non-invasive, low cost, readily available, easy to use and for reasons that are discussed in Chapter 7, it is the preferred option for vocal folds movement measurement in the research reported in this thesis.

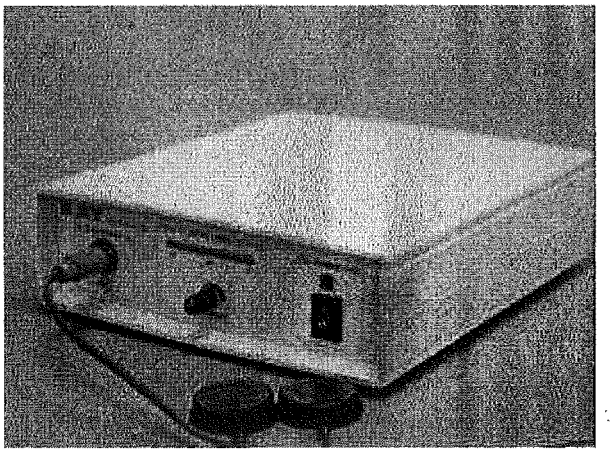

Figure 2.8 EGG device for measuring vocal folds contact measurements (photo adapted from Kay Elemetrics [Ele]).

Chapter 3

Communication techniques for speech impaired individuals

Speech impairment, like any form of physical impairment, happens to people from all walks of life. Although most of us are fortunate enough to be born normal and are able to learn to speak at a young age, there are some who are born mute or speech impaired for various reasons. Not being able to speak prevents these individuals from conveying their thoughts and needs, justifying their actions and so on. Accidents and diseases are the two disasters that occur far too often and are, to an extent, unavoidable. Accident to the neck region from vehicle collision, paralysis from the neck down due to a fall and diseases such as cancer that attack the throat region are some injuries that may cause an individual to become permanently speechless. In some cases, such as for a patient being treated in an intensive care unit (ICU), being unable to speak may only be temporary if the patient recovers well enough to be taken off the ventilator. A study conducted by Bergbom-Engberg and Haljamae [BEH89] with tracheostomised and ventilator-dependent patients reported that the patients' main concern at the time they were unable to talk or communicate was the feelings of anxiety, fear, agony and panic. If help is not available to assist the speech impaired individuals to communicate, they will not be able to express their needs which often lead to anger and frustration for both the caregivers and the patients.

This chapter focuses on the communication techniques for two groups of speech impaired individuals: 1. the tracheostomised and ventilator dependent patients in ICU and 2. the laryngectomees. Each of the techniques described has its own advantages and flaws. The aim is therefore to take advantage of the benefits of the existing techniques and to try to incorporate them in a new design that would allow patients to speak naturally, requiring no or minimal learning experience.

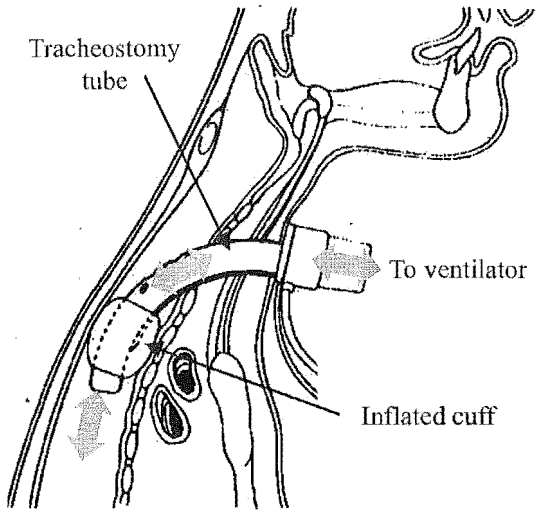

Figure 3.1 Saggital section of the throat region for a tracheostomised patient with an inflated cuff to prevent mucus from flowing into the lungs.

3.1 Speech impairment in tracheostomised and ventilator dependent patients

Tracheotomy is a surgical procedure in which a hole (stoma) is created between the second and fourth tracheal rings and a breathing tube called tracheostomy tube is inserted to allow the patient to breathe through the opening [Bea68]. Tracheal stenosis (constriction of the trachea), tracheomalacia (pathological softening of the trachea) and vocal fold paralysis are some of the conditions that may occur prior to or as a result of the tracheostomy. These conditions may complicate or preclude oral communication [Tip00]. All patients in intensive care unit (ICU) are tracheostomised. Eighty percent of these patients are also ventilator dependent, which means that they cannot breathe on their own and require a machine that is connected to the tracheostomy tube to assist them with their breathing. In such cases, the cuff (a balloon-like structure) on the tracheostomy tube is inflated to prevent secretion from flowing into the lungs. Secretion in the lungs can cause infections that may compromise the patient's condition (see Figure 3.1). With the cuff inflated, the airway is blocked and no air can flow though the vocal folds to cause the folds to vibrate. Therefore, these patients are unable to speak. Although some patients in ICU are kept in induced coma, many are alert but cannot talk because of the tube in place. These patients often get frustrated not being able to express themselves. It has been reported to the author that some of them become so frustrated at not being able to get their needs attended to, they resort to kicking their beds and throwing tantrums at hospital staff.

3.2 Speech impairment in laryngectomees

Laryngectomy is defined as a surgical procedure where the larynx is removed. There are two categories of laryngectomees: partial laryngectomees and total laryngectomees [SS78]. Patients with partial laryngectomy have only part of their larynx removed due to illnesses such as cancer. Moore [Moo75] reported that the portion of the larynx above the vocal folds may be removed with little effect on the voice, but removal of either side of the larynx may cause hoarseness and reduction in voice intensity.

In the case of total laryngectomy, the patient is totally speechless since the whole voice box (larynx) is surgically removed and breathing is usually through a stoma in the throat. The main reason for total laryngectomy is to remove all the structures surrounding the potentially harmful cancerous cells when laryngeal cancer is in an advanced state. The whole laryngeal mechanism, including the epiglottis, hyoid bone, the upper two or three rings of the trachea and the surrounding muscles have to be removed [Soc95, Sal86a, Sal86b, ML90]. In some rare cases, the larynx may also be removed due to traumatic injury to the neck or severe stenosis of the airway [Mar94a].

Unlike ICU patients, laryngectomees are usually alert and can carry on their normal routines in life - except the ability to speak. Without their voice, they lose their sense of belonging and an important aspect of their identity. The focus is therefore on the restoration of speech for total laryngectomees: those with the severe form of speech impairment.

3.3 Standard communication techniques for speech impaired individuals

Over the years, a number of communication techniques have been designed to help improve the lives of speech impaired individuals. These techniques are divided into two main categories: vocal and non-vocal treatments. The non-vocal treatment is further divided into aided and unaided communication techniques. A number of these communication techniques can be used by both ICU patients and laryngectomees. Others may only be suitable for one group of individuals and not the other. The following subsections describe some of the standard communication techniques currently available.

3.3.1 Non-vocal treatment: unaided communication techniques

Unaided non-oral communication is mainly based on gestures, a natural means of communication. Patients use parts of their body such as hand(s) and face, head, eyes, sign language, mouthing the words (lip-reading) and so on to convey a message [Mas93]. For example, a patient may point to his/her mouth to indicate that he/she is thirsty or frown to indicate that something is bothering the patient. Another common technique involves prompting the patient with yes/no questions. The patient replies by nodding/shaking the head or using facial gestures like blinking once to signal

affirmation and twice to signal negation to a given question. The patient may also squeeze the caregiver's hand to indicate a 'yes' to a question and do nothing if the answer is 'no'. Amerind is a formal gestural communication system based upon American Indian hand talk [Ske79]. It is highly predictable and easy to learn.

Sign language, a common non-vocal communication technique used for individuals who are deaf from birth may also be used for laryngectomees and tracheostomised patients. However, sign language is only suitable for a person who already knows the language beforehand. Otherwise, the learning curve may be too steep. For this reason, sign language is not suitable for short-term tracheostomised patients. Lip-reading is another well used method. The patient mouths the words and the caregiver tries to work out what the patient says. Having said that, even trained lip-readers are only able to lip-read with 30-35% accuracy [Bau00]. Ordinary people are rarely able to make out any words beyond very simple sentences.

The non-vocal unaided communication techniques may be the most economical method but they have their downside. The hand gesture and sign language methods for instance require adequate motor skills to be useful, but most ICU patient are too weak to have good hand coordination. The other methods require prompting from an external source. The caregiver has to prompt the patient with the right questions in order to work out the message, just like in a guessing game and this takes time. Patients and caregivers may get impatient quickly if their message does not get through in a reasonably short amount of time.

3.3.2 Non-vocal treatment: aided communication techniques

In the non-vocal aided communication technique, an external object is used to act as a medium of communication between the patient and the care giver. This object can be in a form of charts/boards (see Figures 3.2 and 3.3), pen and paper, Magna Doodle, Magic Slate or a computer [Mas93]. The boards include straightforward alphanumeric characters, pictures and short phrases, sometimes in combination (see Figures 3.2 and 3.3). The caregiver points to a certain part of the chart and if the symbol or phrase on the chart is what the patient had in mind the patient indicates that by nodding the head. This method is a step better than the yes/no questions technique as it sorts through the possibilities in a systematic manner. Patients who use the chart/board method are usually too weak to hold a pen and can only use subtle body movements to indicate yes/no. The user also needs to have good eye-sight.

In the pen and paper, Magna Doodle and Magic Slate option, the patient writes down what he/she has in mind or draws a picture if he/she is illiterate. However this option requires good hand coordination and there is an issue of lack of privacy. A more modern form of the 'pen and paper' method is to use a computer. These days, the user can either type or draw on the computer, save what they have written and use the recorded passages again when needed. This method also needs the user to have

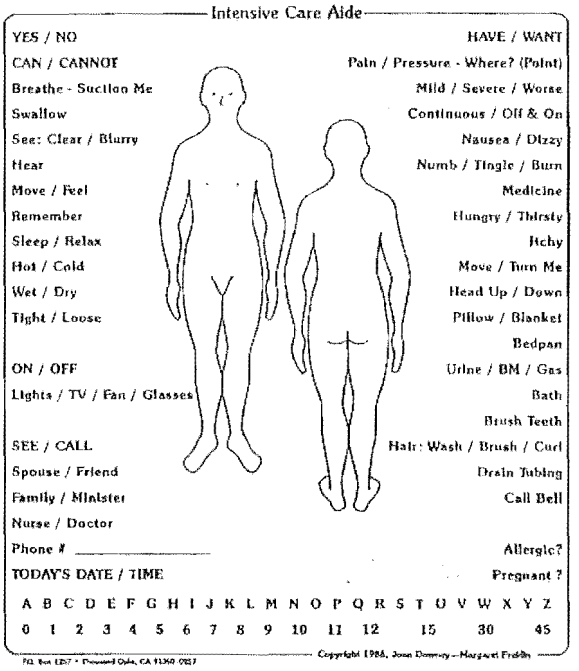

Figure 6-6 The Silent Speaker. (Courtesy of Trademark Corp., Fenton, MO.)

Figure 3.2 Example of an intensive care aid [Mas93].

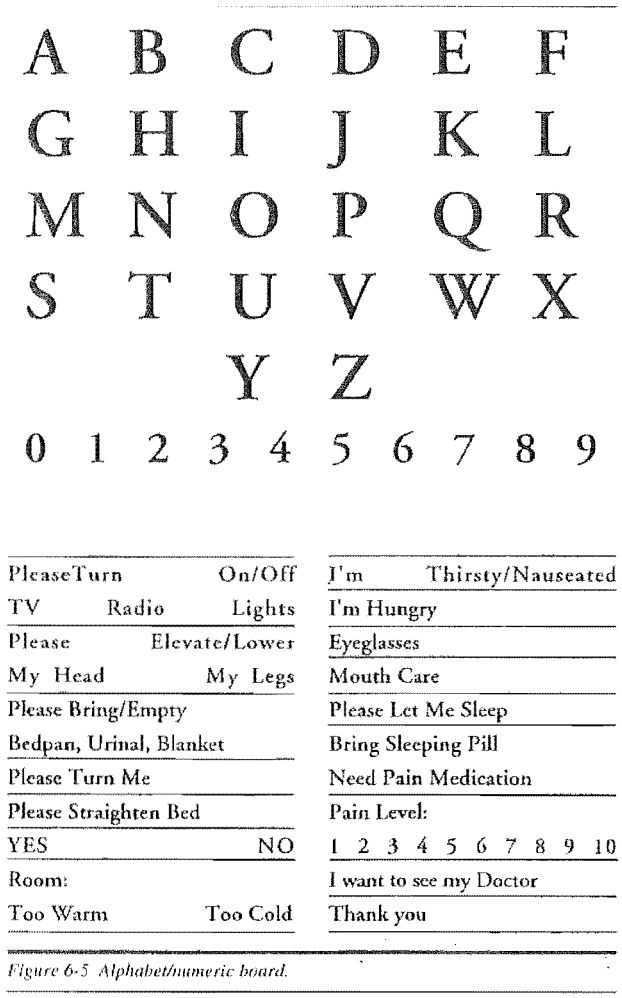

Figure 6-5 Alphabet/numeric board.

Figure 3.3 Example of an alphanumeric board [Mas93].

good eye-sight and to be able to type or draw well. Otherwise, it takes too long to send a message across to the other person.

It has been reported that although some patients find the non-vocal aided method to be quite useful, some complain that it is difficult to read the boards or read what they have written and that they do not have the mental stamina to spell out the words [Mas93].

3.3.3 Vocal treatment

The vocal treatment option for speech impaired individuals relies on the existence of a glottal sound source, supplied by artificial means. It can be in the form of pneumatic devices (if the vocal folds

are intact) or electromechanical vibrating sources. In either case, speech is produced but the quality varies according to the type of device used and also on the patient's physical condition. Vocal communication can be divided into 10 categories: Plugs or buttons method [Tip00], speaking valves, fenestrated tracheostomy tubes, talking tracheostomy tubes, tube-free tracheostomy, esophageal speech, tracheoesophageal speech, pneumatic speech aids, electrical speech aids [Mas93, JJ98] and augmentative and alternative communication (AAC) devices [Ser02].

Plugs or buttons method

This works by plugging the stoma with a stoma plug or button (e.g. Olympic Trach button), diverting the airflow towards the vocal tract instead of the stoma. Stoma plugs are designed for patients who have been taken off the ventilator and that all their speech mechanisms are functional normally. The stomal tract is preserved in case the patient requires a subsequent tracheostomy [Tip00]. They are not suitable for ICU patients on a ventilator or laryngectomees.

Speaking valve

One option that is suitable for tracheostomised patients is the speaking valve. For patients with a simple tracheostomy, air bypasses the vocal tract and is inhaled and exhaled through the tracheostomy tube. By fitting the speaking valve onto the opening of the tube, the air flows into the lungs as the valve opens during inhalation phase. During exhalation phase, the valve closes and airflow is redirected upwards, through the vocal folds and vocal tract, thus allowing the natural production of speech. Like the previous method, speaking valves are not suitable for laryngectomees as the vocal folds have to be intact.

Olympic Trach Talk, Montgomery speaking valve, Hood speaking valve, Kistner one-way valve and Passy-Muir tracheostomy speaking valves are some of the speaking valves currently available [Mas93]. The Passy-Muir valve is the only speaking valve suitable for ventilator dependent patients and the only one that is FDA registered [Mas93]. However, like the other speaking valves, the Passy-Muir is not suitable for inflated cuffed tracheostomy and foam-filled tracheostomy tubes as patients need to be able to exhale on their own to use these valves to speak. Most tracheostomised patients in ICU who are on ventilator are unable to exhale on their own. Another potential problem with the speaking valve is that the valve may pop off when the patient coughs.

Fenestrated tracheostomy tube

A fenestrated tube is a tracheostomy tube that has a single hole or multiple holes on the tube's cannula that allows air to be directed past the vocal folds and through the mouth and/or nose during exhalation (Figure 3.4). They were originally designed for patients who are going through decannulation (removal of tracheostomy tube) process and are not suitable for patients who are ventilator dependent. Shiley fenestrated cuffed tracheostomy tubes and Tucker inner cannula tracheostomy

Figure 3.4 Fenestrated cuffed tracheostomy tube (Mallinckrodt Medical TPI, Inc. Irvine, CA).

tubes are two examples of fenestrated tracheostomy tubes.

The use of fenestrated tubes for speech lead to a number of complications including secretion blocking the fenestration [Mas93, Tip00], growth or granulation tissues in and around the fenestration and occlusion of the fenestration due to tracheomalacia or positioning of tube. All these complications interfere with the speech production process.

Talking tracheostomy tubes

Talking tracheostomy tubes are modified versions of the standard cuffed tracheostomy tubes, designed mainly for patients who are on ventilators and require the cuff to be inflated at all times [Mas93]. It has a separate compressed air supply for speech production that is independent of the respiratory cycle. When the opening of the speech-production-air-supply connector (located outside the tracheostomy tube) is occluded, the compressed air flows through a small conduit that runs along the side of the outer canula and terminates just above the inflated cuff. As the pressure between the cuff and the vocal folds builds up, it forces the vocal folds to vibrate, creating voice as it moves its way through the vocal tract. Some examples of talking tracheostomy tubes are: COMMUNI*trach*^{TM} I, Protex "Talk" tracheostomy tubes (Figure 3.5), foam-filled talking tracheostomy tubes and Bivona talking trach tubes.

A number of potential problems that may be encountered with these devices include the difficulty of the device's placement, patients or caregivers need to manually occlude the air supply connecter in

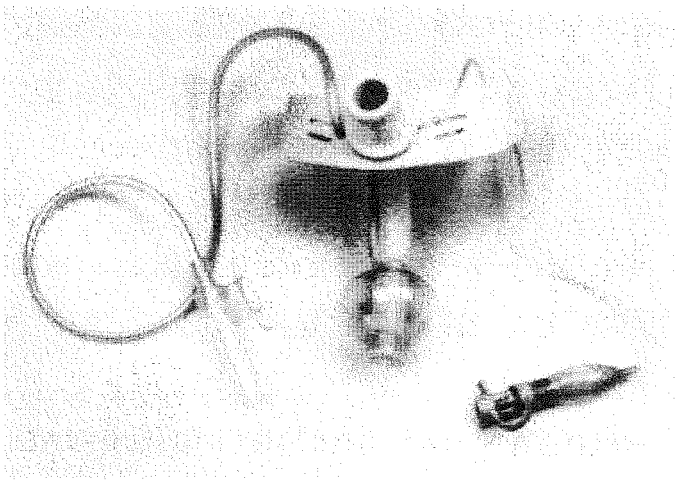

Figure 3.5 Trach-talk, the Protex "talk" tracheostomy tube. (Concord/Protex, division of Smiths Industries Medical Systems, Keene, NH.).

order for the patient to speak and the quality of speech may be quite poor (e.g. gurgle-like, especially when the secretion builds up above the cuff and covers the airflow).

Tube-free tracheostomy

Tube-free tracheostomy is a unique surgical technique in which a stoma is created without a tracheostomy tube in place. It is performed by suturing sections of muscles along the tracheostomy to the side of the neck to form a wall around the stoma [EM03]. The stoma is allowed to heal to form an opening in the trachea.

In order to speak, the patient inhales through the stoma, then constrict the muscles around the stoma to close the stoma and exhale through the mouth and nose. Tube-free tracheostomy apparently works quite well for some patients, especially those who have recovered from ICU and whose stomal tract are preserved in case another tracheostomy is required. Other patients who may be candidates for tube-free tracheostomy are those who has severe airway obstruction, like chronic lung disease, stroke, obstructive sleep apnea, laryngeal stenosis and laryngeal paralysis. Voicing using tube-free tracheostomy requires hours of post-operation practice and cannot be used in conjunction with ventilator dependent patients or laryngectomees.

Esophageal speech

Esophageal speech is an oral communication technique primarily used by laryngectomees. In total laryngectomy, when the larynx and its surrounding structures are removed, the pharynx and esophagus are joined together to form an area known as the pharyngoesophageal (PE) segment [Die91, Ede83]. The PE segment becomes the "new" vibrator (as opposed to the vocal folds) in

esophageal speech [Ede83, DY66, Mar94c].

There are two main techniques for esophageal speech production: the injection and inhalation methods. The injection method is achieved by using the lip muscle, tongue, palates and pharynx to force air into the esophagus, past the PE segment and trapping it for phonation [Kei74, Gra97, Doy94]. As the trapped air returns from the esophagus to the oral cavity, it vibrates the PE segment to create voice [Gra97]. For the inhalation method, the patient takes a quick breath through the stoma, the difference in air pressure between the oral cavity (above the PE segment) and the esophagus (below the PE segment) causes air to be sucked into the esophagus [Kei74, Gra97]. Esophageal speech is produced when air passes through the PE segment upon exhalation. For more detailed information regarding esophageal speech, refer to "The clinician's guide to alaryngeal speech therapy" by Minnie S. Graham [Gra97].

Although esophageal speech does not require any device to operate and is hands-free, it is reported that of all the laryngectomees who attempted esophageal speech, between 40% and 74% (with an average of around 60%) fail to acquire functional esophageal speech [GH61, GRJ+82, KFP68, Mar94c, Put61, Sal83]. Furthermore, the pitch produced in esophageal speech is around 65Hz, substantially lower than that of a normal person's voice (approximately an octave below a typical male voice and two octaves below the female counterpart) [Doy94, Mar94b, RFBS84].

Tracheoesophageal speech

Tracheoesophageal speech is another oral communication technique used by laryngectomees. Surgery is required to create an opening between the esophagus and trachea (known as a fistula) below the pharyngoesophageal (PE) segment. A tracheoesophageal prosthesis is fitted onto the fistula to allow air to flow into the esophagus. The tracheoesophageal prosthesis is a hollow tube with one-way valve that allows air to pass from the lungs into the esophagus but prevents aspiration fluids from esophagus flowing into the lungs[Doy94, CC93, Bos94, Rob94]. The types of tracheoesophageal prosthesis include Blom-Singer voice restoration prostheses, Bivona Ultra Low Resistance voice prosthesis and PROVOX Low Resistance voice prosthesis.

The patient breathes in through the stoma and then covers the stoma with the thumb or a tracheostoma valve (one-way valve similar to the speaking valve used for tracheostomised patients) [Doy94, CC93]. Voicing is produced when the patient exhales and air from the lungs flow though the fistula via the prothesis, past the PE segment, causing the PE segment to vibrate. As the vibrating source flows through the oral cavity, it is modulated into speech by the articulators.

Tracheoesophageal speech is relatively easy to learn and has better speech quality than esophageal speech [Bos94]. Tracheoesophageal speech has also been found to be suitable for patients who have difficulty learning esophageal speech [Gil94, SB80, SBH81]. The drawbacks for tracheoesophageal speech include the need for good manual dexterity and risk of many medical complications such

as *candida* (yeast-like fungi), pneumonia, esophageal stenosis gastroesophageal reflux and so on [Bos94, Gil94].

Pneumatic speech aids

A pneumatic speech aids consist of a vibrator housing with a rubber or steel cup for covering the stoma at one end and a flexible tube that leads to the oral cavity at the other end [Kei74, Sal83, Ler91]. During speech, the device is held against the stoma. When air is exhaled through the stoma it vibrates the housing and sound is produced. Pitch can be varied by changing the air pressure that comes out from stoma [Sal83]. Examples of pneumatic devices are Tokyo artificual larynx and Neher artificial larynx.

Good hand coordination is required when using the pneumatic speech aid as the user needs to periodically lift the cup slightly to allow inhalation and then reposition it again for speech [Blo78]. There may also be problems with the cup leaking that will interfere with the speech produced if the cup is not held properly.

Electrical speech aids

The Electrolarynx, originally developed for laryngectomees, is designed to provide an artificial vibratory source for those whose vocal folds no longer function. There are two types of electrolarynx: neck placement and intraoral systems. The neck placement system (Figure 3.6, left) can be placed against the cheeks, above the larynx (but slightly towards the side of the neck) or the soft fleshy area on the underside of the chin. The problem with the neck placement system is it may be difficult for the patients to find the correct position to place the device as scar tissue on the neck may have left them with little sensation on the neck [Tip00]. Voice intensity is reduced significantly and can be drowned by the vibrator or surrounding sound if a non-optimum position is used.

The intraoral system is similar to the neck placement system but with an extra adaptor that has a tube at one end (Figure 3.6, right). The vibration from the electrolarynx is transferred through the tube into the patient's oral cavity. Speech is achieved by mouthing the words. For individuals who have a fuller figure, the electrolarynx may work better if an intraoral system is used or if the neck placement system is placed against the cheek, as the vibrating signal has to be able to couple through to the vocal tract. It has been reported that four out of five temporary tracheostomised patients were able to use the electrolarynx to communicate with their caregivers [Sun73].

Earlier forms of electrolarynx like the ones shown in Figure 3.6 are monotonous. The nature of the signal makes the speech sound very mechanical (robot-like). An improved version of the electrolarynx is the Trutone [Gra97]. Trutone has all the functions of an electrolarynx plus the option of manual pitch variation through a pressure sensitive switch. It requires a conscious effort from the user to vary the pitch as the user speaks. To use the electrolarynx or Trutone, patients have to have

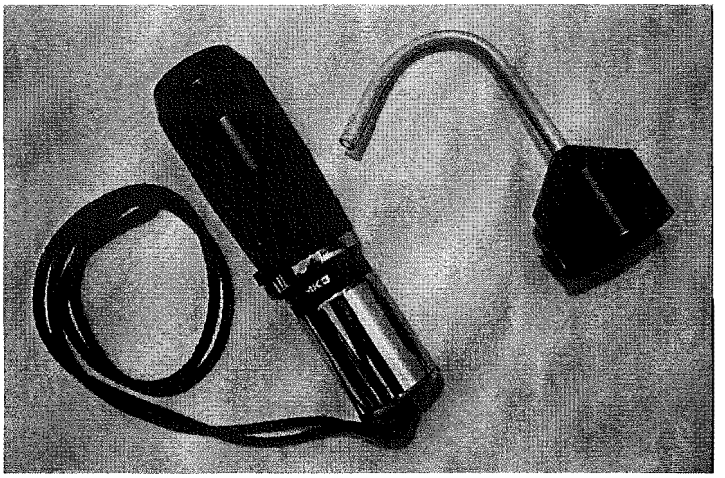

Figure 3.6 Neck type electrolarynx (left) and the intraoral adapter (right).

good hand control (to hold the device in hand) and hand-speech coordination (to know when to vary the pitch so that it sounds natural).

Yet another improvement from the Trutone is the UltraVoice [Ult02, HSLDK95]. It was designed for laryngectomees but may also be suitable for tracheostomised and ventilator dependent patients who have reasonably good motor skills. UltraVoice is consists of an oral unit and a control unit. The oral unit comes in a denture form or a retainer form that can be placed inside the mouth. Inside the oral unit are a radio circuit, rechargeable batteries and a speaker. During speech production, the user switch on the device and the control unit sends a signal to the oral unit. The control circuit demodulates this signal and sends it out as glottal pulse through the speaker. The user then speaks by mouthing the words. The UltraVoice has three modes of speech: male voice, female voice and whisper.

The advantage of the UltraVoice is that it is wireless; the user does not need to operate the device with his/her hand near the mouth. However, it still requires a certain amount of manual control in the on/off switching of the device. Since the vibrating device is in denture/retainer form, it has to be custom made. Hence, it may not be suitable for short term users such as ICU patients.

3.3.4 Alternative augmented communication (AAC) devices

Alternative augmented communication (AAC) devices are electronic devices that can produce speech output [Ser02]. These devices can be used to produce digitised human speech or synthesised text-to-speech conversion. Some devices can be preprogrammed to produce preplanned messages or can be set up for the production of spontaneous novel messages via an on-screen keyboard. There are

also a number of other access methods including the standard keyboard, switch, mouse and joystick. Scanning mode is used to enable patients with limited movement to access the device.

Some of the AAC devices need the user to have good eye-sight (e.g. to see the screen) or to have good hand coordination (to operate a mouse, joystick, switch or type on a keyboard) in order to trigger a predefined message or to produce a spontaneous message. The habitual pitch is usually constrained to that of a fixed male or female voice and therefore not the user's own habitual pitch. The text-to-speech voice is clear but pitch variation is unnatural, causing the speech to sound mechanical. AAC devices may be more suitable for laryngectomees than for ICU patients.

3.4 Proposal for improved speech production for tracheostomised individuals and laryngectomees

Through years of research, scientists and engineers have come up with many ideas of speech aids for the speech impaired individuals, most of which have been covered in this chapter. The ultimate aim is to come up with a device that will replace the missing structure(s) of the speech mechanism so that normal speech can be restored.

All of the speech aids discussed in this chapter have some form of drawback:

- They may require the user to have relatively good motor skills: sign-language, pen and paper, pneumatic speech aids and tracheoesophageal speech, for example, may be suitable for laryngectomees but not tracheostomised and ventilator-dependent individuals as they are often too weak to control these devices.
- They may require good eye-sight: speech charts, typing on the computer, and pen and paper techniques all require the patients to have good eye-sight in order to work properly.
- They may take too much time to learn: some communication techniques, such as tube-free tracheostomy and esophageal speech, require an initial learning stage involving a lot of time and effort from both the patient and the caregiver.
- The voice produced may be unnatural (monotonous or robot-like): the quality of voice generated with artificial speech device (e.g. electrolarynx and Trutone) are monotonous, making the voice sound very unnatural.
- Users may lose the identity their of own voice: the sound source generated from mechanical speech devices does not resemble the shape of the glottal pulse, it is merely the impulse response of a plastic diaphragm being periodically knocked by a spoke. This is one of the reasons that artificial speech aids sound so unnatural.

- It may be too time consuming to convey even simple messages: although speech charts and pen and paper are quite straightforward to use, in terms of application they take up quite a lot of time and concentration from the patients as well as their caregivers.

Therefore, there is a need for a device that has minimal learning time, easy to use and capable of generating a natural sounding glottal sound source. This thesis proposes an artificial speech device that is designed to overcome some of the problems mentioned above. Some features included in the new design are:

- Minimal learning requirement: the user mouths the words and the device generates the appropriate voice. It is similar in principle to the electrolarynx but has additional features to improve the quality of the artificial speech.
- Hands-free operation: 1. the user can 'wear' the device with its specially designed headset instead of having to hold it every time the device is used and 2. the switching of the sound source can be done automatically as opposed to manually flipping a switch to turn the device on/off (e.g. the UltraVoice in the previous section).
- Automatic pitch control: the device tracks the motion of the jaw to automatically control pitch. Pitch variation is one of the most important components of natural sounding speech.
- More natural sounding voice with the options of shifting the habitual pitch: a model of glottal waveform that varies with pitch is required to produce a sound source that approximates the actual glottal source signal.

Chapter 4

EGG analysis techniques

In this chapter, the EGG signal is introduced as an alternative to glottal airflow as the glottal sound source of an artificial speech device. A detailed explanation for the use of EGG as a glottal sound source can be found in Chapter 7.

The electroglottograph (EGG) is an instrument which indirectly measures the vocal folds contact by measuring the impedance between two electrodes placed externally on either side of the thyroid cartilage. As the vocal folds come together during voicing, the impedance between the electrodes decreases; when the vocal folds move apart, the impedance increases [CHMA86]. The EGG device is relatively low cost, non-invasive and easy to operate. Since the opening duration of the vocal folds is usually longer than the closing duration, the EGG waveform has the shape of a skewed sinusoid (see Figure 4.1, bottom). This EGG signal also has an 8-12dB/oct slope in its spectrum, similar to that of glottal airflow. It is also independent of the vocal tract shape (e.g. vowel independent), which makes the design of the glottal model simpler.

Rosenberg [Ros71] and Liljencrants and Fant [FLL85] presented two glottal sound source models that are often used for voice synthesis. These models were based on glottal airflow waveforms derived from inverse filtered acoustic signals. Glottal airflow has an 8-12 dB/Oct slope in its spectrum. An example of glottal airflow signal is shown in Figure 4.1, top.

There are a number of techniques for analysing glottal waveforms but there are no rules as to which method is the best. Although there are some analysis techniques that are more commonly used than others, different investigators use different methods to cater for their needs. The main difference between the various methods lies in the setting of the threshold values for determining the different phases of the vocal folds movement. There are four separate sections in each cycle of the glottal waveform: the opening phase (*OP*), closing phase (*CP*), opened phase (*Open*) and closed phase (*Closed*). Assuming that the vocal folds are adducted at the start, *OP* is the interval over which the

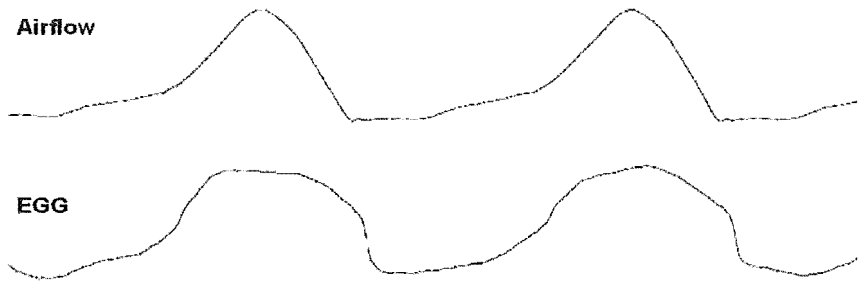

Figure 4.1 The figure above shows an example of the the airflow signal (top) and EGG signal (bottom) [Pul05].

vocal folds start to separate (e.g. when vocal folds abduction occurs). As the folds separate, they lose contact. The interval over which contact is lost is defined as *Open*. At some point in time, the vocal folds start to come together during an interval defined as *CP*. *Closed* is the interval where the vocal folds are fully adducted.

In this research, the EGG signal is used instead of the glottal airflow signal to represent the glottal source. Although the *OP* and *CP* in EGG may not always correspond to the instant of glottal opening and closing respectively [BO00], it is not important in this situation because the purpose for obtaining these parameters is to provide strategic points on the waveform that can be used as inputs for glottal waveform modelling (see Chapter 5). There is no intention to associate the *OQ* and *SQ* measures obtained from the EGG signals with the corresponding *OQ* and *SQ* parameters obtained from airflow signals. The measures will however be useful in the instance where the EGG signal is used as glottal source for artificial voice simulation (Chapter 7).

The voice source, for the purpose of this research, is characterised by the average *F0*, *OP*, *CP*, slope at the closing phase (*sCP*), open quotient (*OQ*) and speed quotient (*SQ*) obtained from the EGG and differentiated EGG (*DEGG*) signals. In the following sections, methods for analysing the EGG signals and extracting the parameter values are reviewed. The EGG parameters extraction techniques used in this thesis are discussed in Section 4.4.

4.1 Threshold method

One of the most common methods used for analysing the EGG waveform is to set the lower and upper threshold values of the signal at 10% and 90% levels of every cycle and to analyse each cycle of the signal separately [Mar96] (see Figure 4.2). The interval over which the signal lies below the 10% level is defined as *Closed*. The interval over which the signal exceeds the 90% level is defined as *Open*. The remainder of the EGG cycle comprises the *OP* and *CP* intervals as shown in Figure

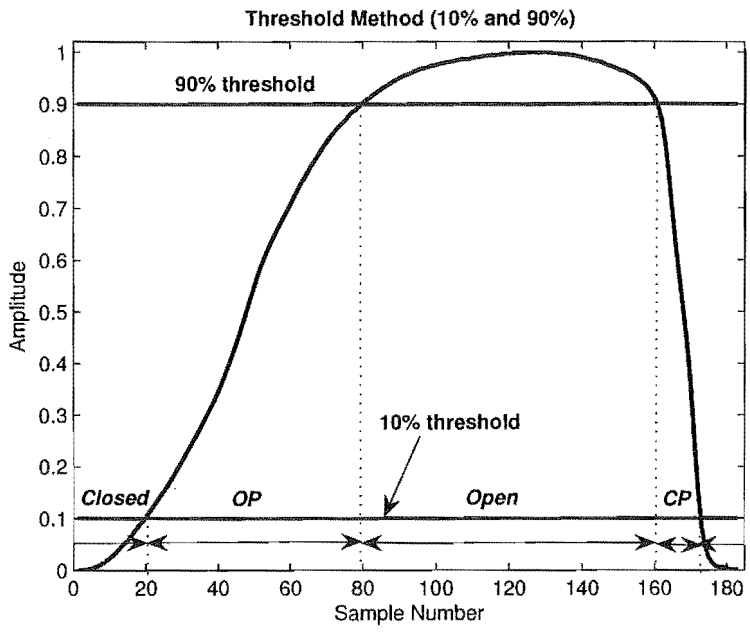

Figure 4.2 The double threshold method for EGG waveform analysis.

4.2.

Figure 4.3 shows another method where only 1 threshold (90%) is used. Only three phases are then defined for the EGG cycle as shown in Figure 4.3.

The parameters are extracted from each cycle of the EGG waveform. The mean values of the parameters over several cycles are then calculated and used for comparing the results of different waveforms.

4.2 LF-Model method

The next method is the LF-Model (Liljencrants and Fant) technique [FLL85]. This method is used for both waveform analysis and waveform synthesis. It is a four-parameter model commonly used for of glottal airflow signals (see Figure 4.4, top). In this case, the glottal parameters are obtained by means of the slope of the signal (Figure 4.4, bottom). The *OP* begins at the start of the pulse when the slope is zero to when the slope reaches its positive peak. *Open* is the duration between the positive peak and the negative peak. The negative peak indicates the instant where the vocal folds are closed. *CP* is the time between the negative peak and the time when the differentiated signal reaches zero. *Closed* is the interval during which the slope of the signal remains at zero. As for the

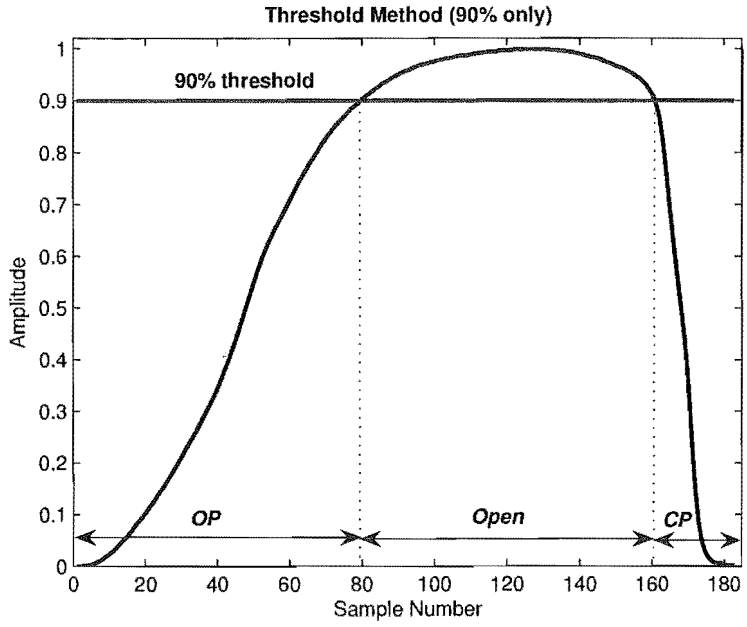

Figure 4.3 The single threshold method for EGG waveform analysis.

threshold method, each pulse is analysed separately.

4.3 Computerised Speech Lab (CSL) and Speech Station2

Computerised Speech Lab (CSL) [sl00] and Speech Station2 [Sen] are commercial audio processing packages for analysing speech signals and for other functions such as data acquisition, graphical and numerical displays, and signal editing. These software packages are good for finding pitch, jitter (cycle-to-cycle pitch variation) and shimmer (cycle-to-cycle amplitude variation) from acoustic signals, but there are a few parameters required in EGG analysis (e.g. *OP* and *CP*) that are not available from these packages. Therefore, a more specific program was developed by the author to include the measurement and calculation of these parameters. The algorithm for this software is described in the next section.

4.4 Modified EGG analysis technique

The EGG analysis technique used in this thesis is based on the LF model with a few minor alterations. In this method, *Closed* is included in *CP* since the closed phase of the EGG signals as observed from all the experiments conducted appears to be quite small (less than 2% of the total

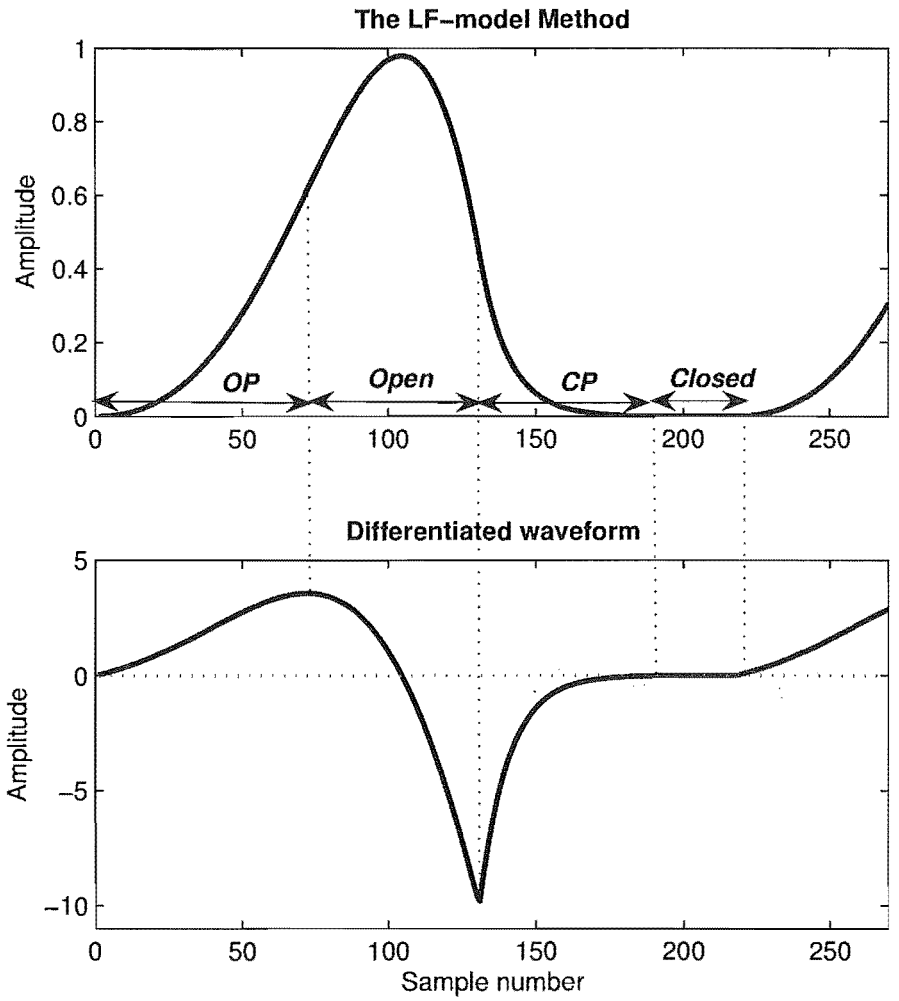

Figure 4.4 The LF-model method.

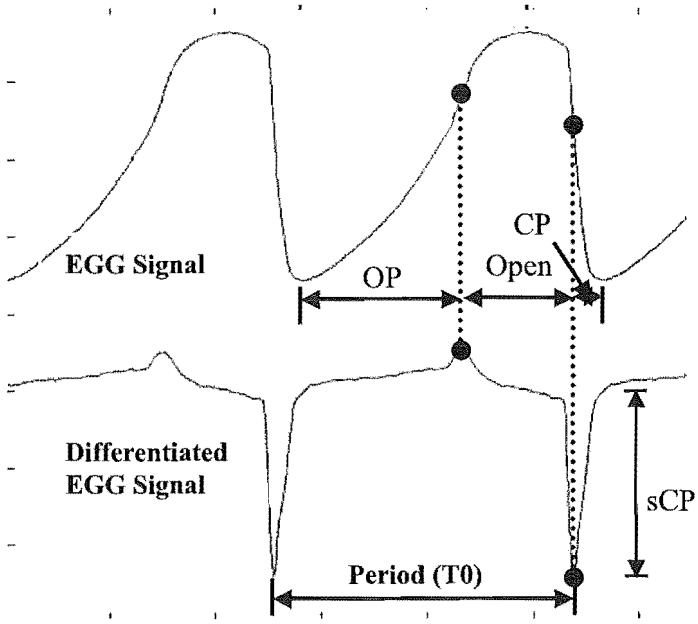

Figure 4.5 Modified EGG analysis method. An example of an inverted EGG waveform and its time derivative waveform showing how *T0*, *OP*, *CP*, *Open* and *sCP* are obtained.

pulse length). The other difference is that the average waveform is calculated before the signal is analysed. This gives a better indication on the average shape of the glottal pulse, which is important when the aim of the project is to generate a synthetic glottal pulse that resembles the original human glottal pulse.

The OQ and SQ values are two quantities that are commonly used by speech scientists to describe the shape of a particular glottal waveform. OQ and SQ are not directly measured from the EGG waveform but derived from Eqns 4.1 and 4.2 respectively.

$$OpenedQuotient,\ OQ = \frac{Open}{T0} \tag{4.1}$$

$$SpeedQuotient,\ SQ = \frac{OP}{CP} \tag{4.2}$$

where OP begins at the minimum point of the EGG waveform and ends at the point that corresponds to the maximum point on the DEGG waveform. CP begins at the point that corresponds to the minimum point of the DEGG waveform and ends at the next minimum point on the EGG waveform. *Open* is the duration between the maximum and minimum points on the DEGG waveform (see Figure 4.5). *T0* is the period of the waveform. Henrich *et al* [HRC03] used this same technique to

find the *OQ* values for their experiment.

The following sub-sections describe the EGG waveform analysis algorithm written specifically for all the experiments carried out in this thesis. This algorithm was written to analyse EGG waveforms in .wav file format.

4.4.1 Signal segmentation

The signal analysis software from MATLAB's signal processing toolbox (SPTOOL) was used to assist the experimenter in isolating the section of signal to be analysed (see Figure 4.6). As audio and EGG signals were recorded simultaneously in all experiments, the use of MATLAB toolbox software enabled both audio and EGG waveforms to be displayed on the screen, allowing sections of the waveforms to be played and listened to. The segmentation process was done manually based upon a number of criteria:

- The audio signal was listened to and compared with the EGG signal during the selection process to ensure that the segment of EGG selected corresponded to the vowel of interest.
- The segment selected had to be in a region where the signal was stable. This region is usually in the middle of an utterance where the amplitude and *F0* of the signal were relatively constant.
- Signals that were too weak (low SNR) were discarded as noisy averaged waveforms reduce the accuracy of the parameters extracted from the waveform analysis process.
- Signals with modal mode (single peak per cycle) and vocal fry mode (multiple peaks per cycle) were analysed separately to avoid the distortion of the averaged waveform.

The actual signal analysis was carried out by the EGG_Waveform_Analysis software (see Figure 4.7 for the flow diagram of the software) written in MATLAB specifically for analysing EGG waveforms.

4.4.2 Drift removal

After the signal of interest had been selected, the low frequency signal known as drift was removed otherwise the average EGG signal would be distorted. This low frequency artifact was largely contributed by the movement of the subject's body during the recording process. It did not affect the overall quality of the sound. Two EGG signals, one the original and the other with the drift signal removed were played to an audience comprising two speech language therapists, three engineers and a medical doctor. All parties agreed that the signal sounded the same with or without the drift.

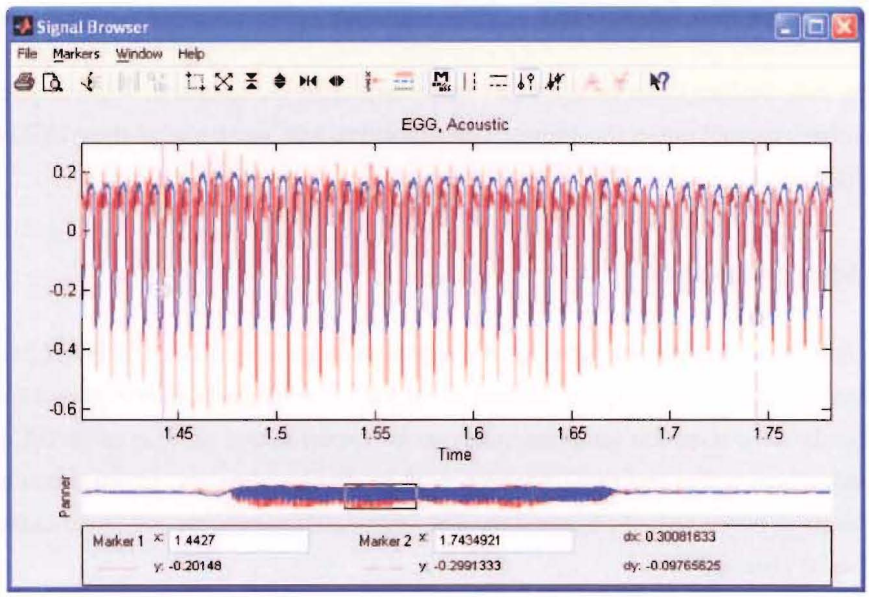

Figure 4.6 Signal segmentation using MATLAB's signal processing toolbox (SPTOOL).

Algorithm for removing drift

The EGG signals recorded using electroglottography (Kay Elemetrics Model 6103) were raw or unprocessed data. There are other commercialised electroglottographs that would pre-process the EGG signal (by means of a high-pass filter) to remove the drift before the signal is digitised. The drawback of removing the drift this way is that it distorts the shape of the signal.

One way of reducing the distortion is to remove the drift cycle-by-cycle. The simplest yet effective method is by locating the start and end point of each EGG cycle, and then removing the baseline shift of the waveform by subtracting from each point on the EGG signal a value equal to the linear interpolation between the two points (as shown in Figure 4.8). The resulting pulse starts and ends at zero amplitude. This process was repeated until the last pulse of the signal was reached.

4.4.3 Signal differentiation and average frequency calculation

With the drift removed, the EGG signal was then differentiated (to determine the slope). For each individual EGG pulse, the maximum slopes for both the rising edge and the falling edge of the signal which corresponded to the positive peak and negative peak of the differentiated signal were located. The duration between two consecutive (positive or negative) peaks indicated the period, $T0$, of the pulse. The preliminary average frequency (habitual pitch, μF_0) was calculated and pulses that were beyond $\pm 2.5\%$ of μF_0 were discarded to minimise distortion to the average waveform.

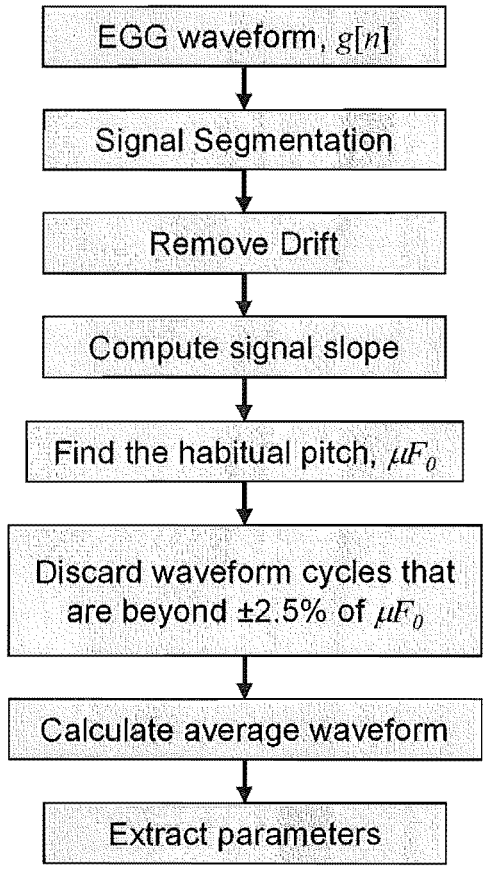

Figure 4.7 EGG signal analysis software (EGG_Waveform_Analysis) flow diagram.

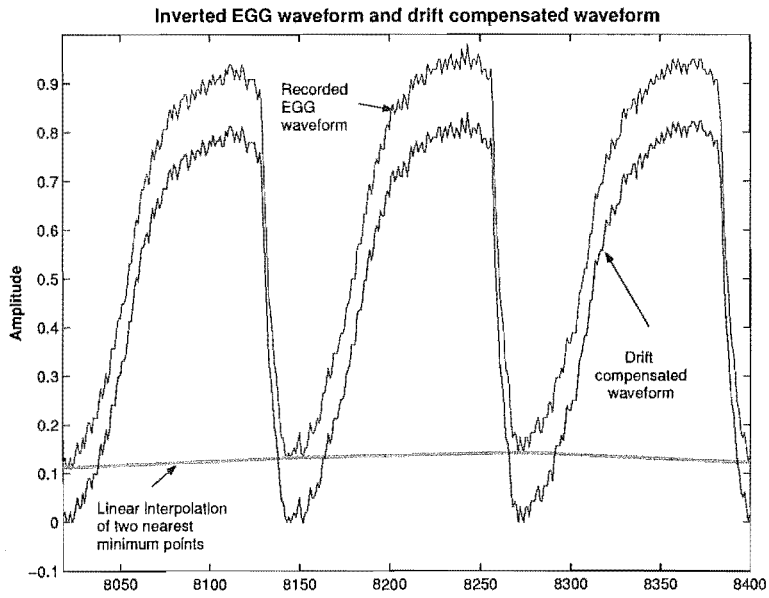

Figure 4.8 Recorded EGG waveform and the same waveform with baseline shift removed.

The DEGG estimation

The DEGG or the slope of the EGG signal, $g'[n]$, was obtained by computing the best-fit straight line to each set of consecutive points, e.g. linear regression. The equation used for calculating the DEGG signal is:

$$g'[n] = \frac{\sum_{m=-\frac{k-1}{2}}^{\frac{k-1}{2}} (m \times g[n+m])}{\sum_{m=-\frac{k-1}{2}}^{\frac{k-1}{2}} |m|^2} \tag{4.3}$$

where k is the number of samples of $g[n]$ used to estimate each derivative estimate (an odd number approximately corresponding to 10% of the average EGG pulse length). Before $g'[n]$ was calculated, the average EGG signal, $g[n]$, was extended by k samples at both ends of the waveform by repeating the pulse. The first and last $\frac{3k-1}{2}$ points on the resultant waveform were then removed to obtain the actual slope, $g'[n]$. Figure 4.9 shows the EGG signal and its estimated slope. To demonstrate the relationship between the EGG signal and the DEGG signal, the slope at particular points on the DEGG signal (illustrated by the dots on the DEGG signal) was calculated using $k = 13$ samples on the EGG signal (represented by the straight lines superimposed on the EGG signal).

To obtain a more accurate estimation of the DEGG signal, an extra weighting variable could be

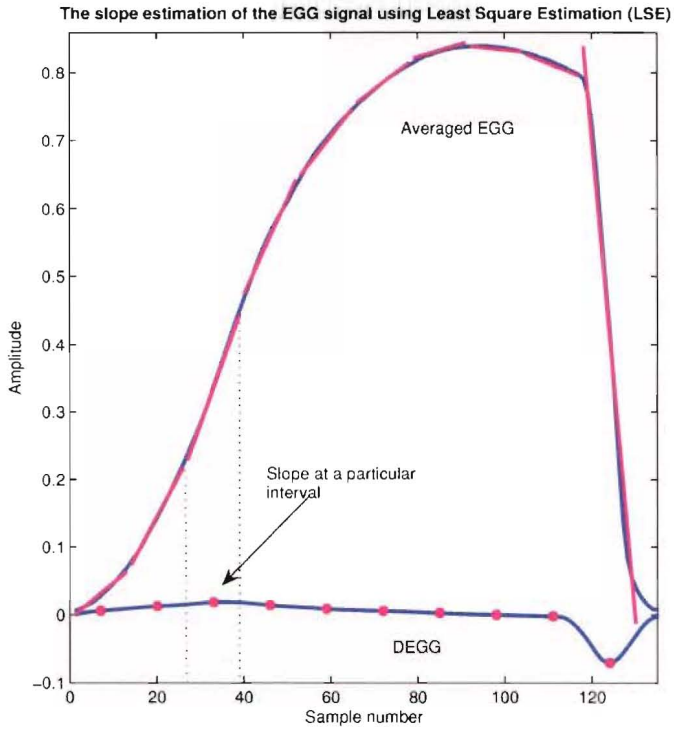

Figure 4.9 An example of the averaged EGG waveform (top) and the slope of the signal, DEGG (bottom).

introduced so that sample points that were closer to the point where the slope was to be calculated had a higher weighting than those further away from the point of interest. However, since the EGG signal was sampled at $f_s = 22050$Hz (sufficiently over-sampled compared to the usual sample rate of 4000Hz) and the signal was relatively low noise, the current version of calculating the slope of the EGG signal was sufficient.

4.4.4 Average waveform calculation

The alignment of pulses is important in estimating a meaningful average EGG waveform. After the habitual pitch, μF_0, had been calculated and pulses that were beyond $\pm 2.5\%$ of μF_0 removed, the remaining pulses were used to calculate the average EGG waveform.

Tests were carried out to find the most suitable method for calculating the average waveform. Four methods were considered, where pulses were aligned at the:

- Start of each pulse
- Peak of each pulse
- End point of each pulse's *OP* (corresponds to the instant of the positive peak of the pulse's DEGG waveform)
- *CP* onset of each pulse (corresponds to the instant of the negative peak of the pulse's DEGG waveform)

Figure 4.10 shows an example of the alignment of three EGG pulses using the four methods. Out of the four methods considered the *CP* onset pulse alignment method resulted in the lowest standard deviation. This was largely due to the fact that the slopes of the pulses along the falling edge were fairly constant (provided that they were in the same voice register) and only varied significantly across individuals, not within an individual. Therefore, this method was employed in finding the average waveform. The next section discusses the *CP* onset pulse alignment method.

Closing phase onset alignment

Each pulse on the EGG signal was located, truncated and then aligned on the onset of *CP*, the steepest point of the pulse's falling edge (see Figure 4.11, left). Pulses that were shorter than the longest pulse were extended by repeating the pulses at both ends of the pulse so that all pulses had the same length as the longest pulse. The average pulse was calculated by taking the average of each point at the same instant in time. The standard deviation was also calculated in the same manner. The beginning and end of the pulse were located as the minimum point to the left and to the right of the peak average pulse respectively. Figure 4.11, right, shows the average EGG waveform calculated using the *CP* onset alignment method.

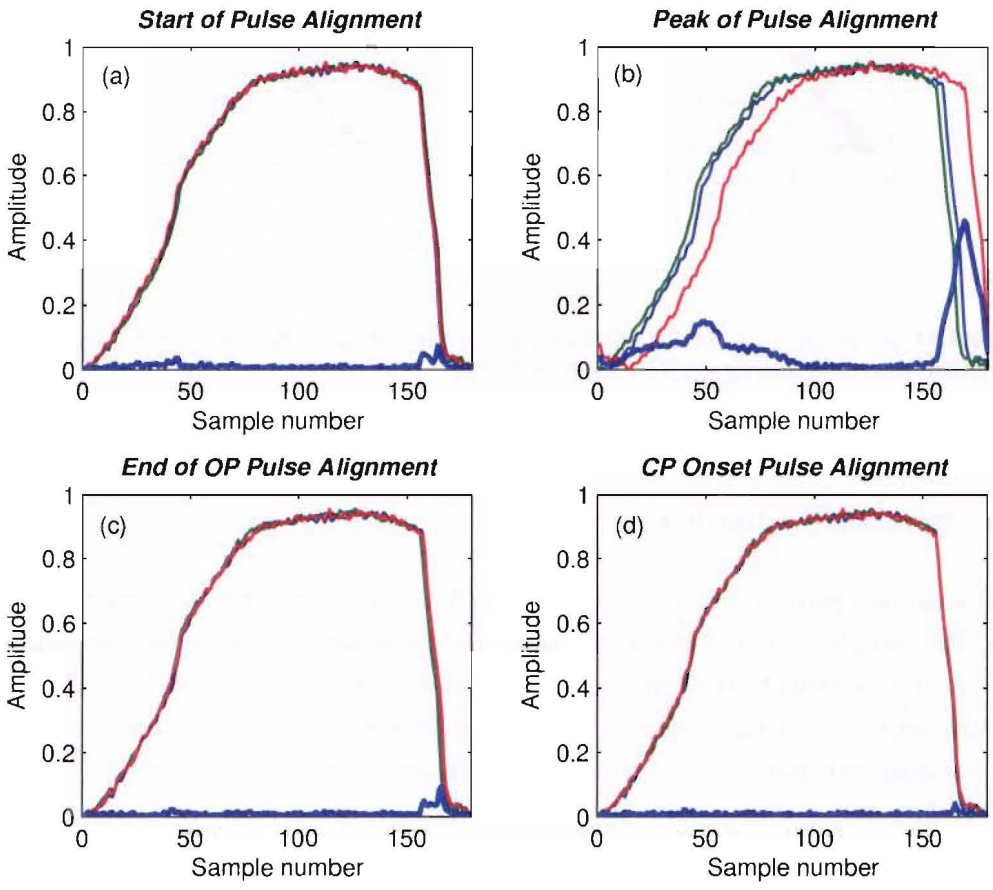

Figure 4.10 The waveform alignment methods: (a) start of pulse alignment, (b) peak of pulse alignment, (c) end of OP pulse alignment and (d) CP onset pulse alignment. Each plot shows the alignment of three EGG pulses and their standard deviation (shown by the thick line at the bottom of each plot).

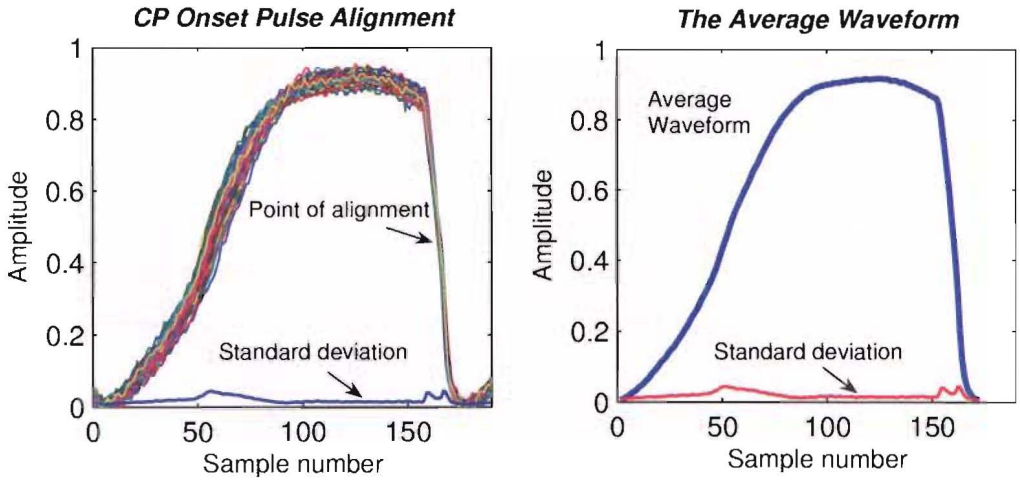

Figure 4.11 An example of the *CP* onset alignment method including the standard deviation curve (left) and its resultant average EGG waveform (right).

4.4.5 Parameters extraction

With the average pulse available, the average glottal parameters can then be measured and calculated. The average pulse was first differentiated and the instants where the positive and negative peaks occur were located. As before, the averaged *F0* was obtained through the inverse of the average pulse length ($\frac{1}{T0}$). *SQ* and *OQ* were then derived from these parameters. These two parameters were important indicators of pulse shape. *SQ* determined the skewness of the pulse while *OQ* the width of the pulse.

Besides the modified LF method, two other methods were also included in the signal analysis. They were the second differential technique and the 90% threshold method. The second differential technique was exactly the same as the modified LF method, except that the positive (*OP*) and negative (*CP*) peaks were located based on the second derivative of the average EGG pulse rather than the first derivative. For the 90% threshold method, the end point of *OP* and the starting point of *CP* were the intersect points between the average pulse and its 90% horizontal threshold line. With *T0*, *OP* and *CP* values, the other parameters (*F0*, *SQ* and *OQ*) were calculated using exactly the same formula as in the modified LF method. An extra parameter later included in the analysis software was the maximum negative slope of the DEGG signal (*sCP*). It was used in an experiment related to the glottal waveform shape (Chapter 7) and also for a new glottal waveform modelling that will be introduced in section 5.3 of this thesis.

It was concluded that of the three techniques of analysing the average EGG pulse, the modified LF

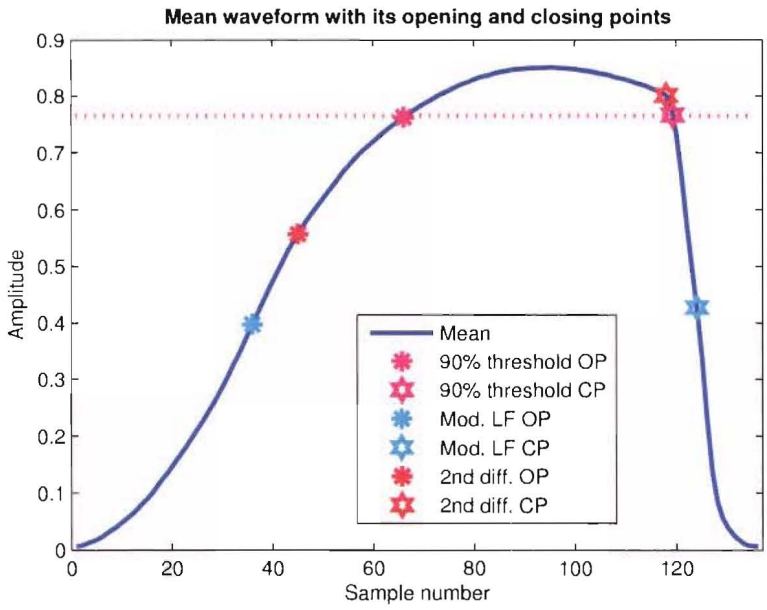

Figure 4.12 The three methods of finding *OP* and *CP* points on the EGG waveform: (a) the 90% threshold method (magenta), (b) the modified LF method (cyan) and (c) the 2^{nd} differential technique (red).

Table 4.1 Results of *SQ* and *OQ* from all three methods (example).

Method	*SQ*	*OQ*
90% threshold	3.88	0.39
Modified LF	3.00	0.65
2^{nd} differential	2.50	0.54

method was the most reliable method as the peaks could be easily identified. See Figure 4.12 for the three techniques and their corresponding *OQ* and *SQ* values in Table 4.1. The second differential technique was less reliable mainly because by differentiating the signal twice, more noise was introduced into the system making it harder to locate the peaks. The 90% threshold method on the other hand used an arbitrary threshold to locate the parameters. The threshold value chosen did not have any physiological relationship with the vocal fold movements. Therefore, it was not surprising that the results yielded from this method were also not very reliable. The modified LF method is used for all of the results of EGG analysis in the thesis.

Chapter 5

Glottal Pulse Model and Vocal Tract Model

This chapter introduces glottal pulse models and vocal tract models for voice synthesis. It begins by presenting the analogy between the speech mechanism for normal speech production and for artificial speech production. A model for the glottal pulse is important because the naturalness of voice depends on the glottal pulse shape [Ros71]. Next, a review of well-known glottal pulse models is presented, followed by the introduction of a new glottal pulse model known as the "twin-bar model". To compare the quality of the twin-bar model with the existing glottal models, synthesised voice using these models is generated. A vocal tract model is needed to provide a filter for the glottal source so that synthesised voice can be generated. The vocal tract model which is based on the waveguide model ensures that the vocal tract shape for a particular vowel is consistent.

5.1 The analogy between natural and artificial speech

When a person speaks, a stream of air is forced from the lungs into the trachea through the vocal folds, causing the vocal folds to vibrate (see Figure 5.1). The positions of the tongue and lips determine the sound that is produced when the vibrated air flows through the vocal tract. In terms of artificial speech, the lungs are analogous to a power supply while the vibrating vocal folds are equivalent to an oscillator and the resultant sound is an artificial voice source. From there, the sound travels through the vocal tract as in normal speech. The goal here is to replace the function of the vocal folds with a specially designed oscillator. Figure 5.1 shows the analogy between speech mechanism for normal speech production and for artificial speech production.

5.2 Existing glottal pulse models

A simple method of generating a glottal pulse is to use the average waveform obtained from EGG recordings of normal subjects with different voice types as templates. Pitch change can then be

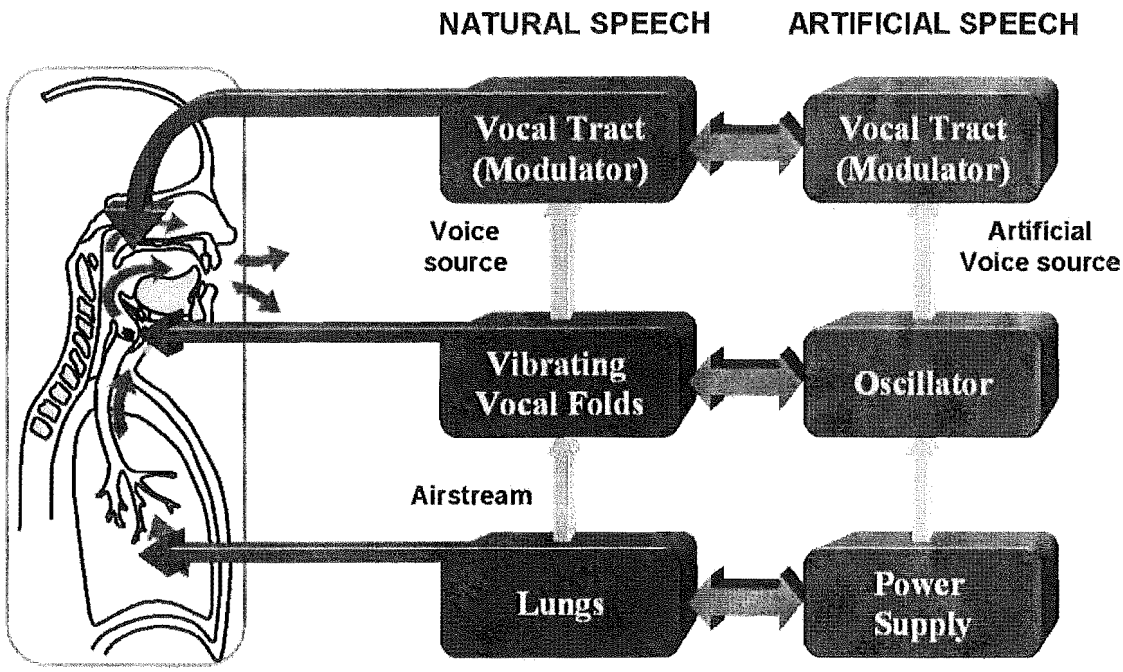

Figure 5.1 The analogy between natural and artificial speech.

modelled with linear expansion and compression of a template via linear interpolation. This way, the overall shape of the waveform is retained. However, in reality, to improve the naturalness of synthesised speech subtle shape changes that vary with pitch have to be taken into account. Hence, a model for the glottal pulse is required.

This section reviews well-known glottal pulse models (e.g. all-pole model [JHP00], Rosenberg model [Ros71, JHP00], LF model [FLL85] and physical models [FL68, IF72, ST95, ABT00]).

5.2.1 Two-pole model

The glottal spectrum of a normal phonation has a slope of approximately -12dB/octave [Fla57]. Since each pole in a system contributes to a 6dB/octave attenuation in the frequency domain, one logical method of generating a glottal pulse is to represent the signal as a two-pole low-pass filter [Fan60], where the glottal pulse is the impulse response of this filter. Eqn. 5.1 shows the expression for the Z-transform of the two-pole filter, $G_{2P}(z)$.

$$G_{2P}(z) = \frac{G_0}{(1-\alpha z^{-1})(1-\beta z^{-1})} \tag{5.1}$$

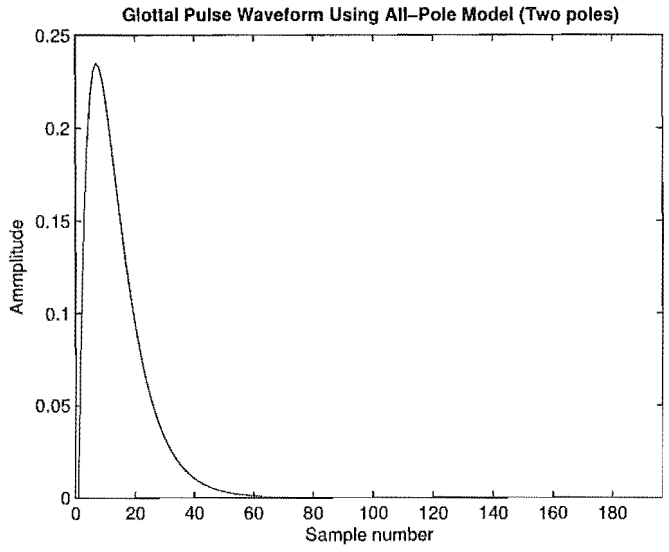

Figure 5.2 The all-pole (two poles) glottal pulse.

The value G_0 is the gain constant, α and β are real poles inside the unit circle where $\beta < \alpha < 1$ and $\alpha \approx 1$.

Usually the time domain expression of the two-pole filter is the preferred option as the input parameters seem more natural and the output produces perceptually better results. Figure 5.2 shows the glottal pulse of the time domain two-pole model. This model can be expressed as:

$$g_{2P}[n] = (\alpha^n - \beta^n)u[n], \tag{5.2}$$

where $u[n]$ is the unit step sequence.

Although the two-pole model is simple to implement, it is not suitable for this research because this model cannot produce pulse shapes with the *OP* duration longer than the *CP* [JHP00, Del83], an important characteristic of a glottal pulse, and the vocal quality is poor [CW90].

5.2.2 Rosenberg model

One of the more popular glottal pulse models is the Rosenberg model. This model was developed in the early seventies. It uses trigonometric functions to form the glottal pulse and has a single point of discontinuity at the end of the open phase. Perceptual tests carried out by Rosenberg found that pulse shape with a single slope discontinuity sounded more natural than pulses with more than one

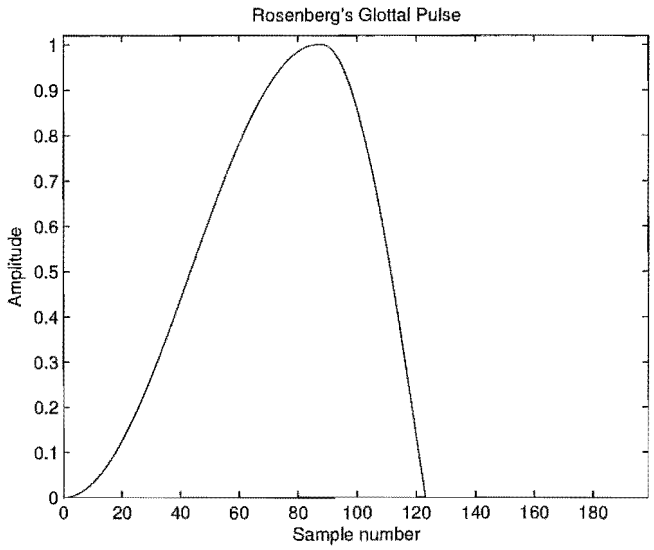

Figure 5.3 Rosenberg's glottal pulse.

slope discontinuity or those with no discontinuity [Ros71]. The following equation (Eqn. 5.3) shows the glottal pulse model by Rosenberg.

$$g_{RO}[n] = \begin{cases} \frac{1}{2}[1 - \cos\frac{\pi n}{N_1}] & for\ 0 \leq n \leq N_1 \\ \cos\frac{\pi(n-N_1)}{2N_2} & for\ N_1 \leq n \leq N_1 + N_2 \\ 0 & otherwise \end{cases} \tag{5.3}$$

where $N = \frac{f_s}{F0}$, $N_1 = 0.4N$ and $N_2 = 0.16N$.

An example of the Rosenberg's glottal pulse is as shown in Figure 5.3. The Rosenberg model is one of the two models (the other being the LF-model) that have been used to compare with the twin-bar model. The quality of artificial voice generated with the 3 different glottal models is discussed in Chapter 7.

5.2.3 LF-model

The LF-model stands for Liljencrants and Fant model, the names of the creators of this model. This model was originally designed for the derivative of the volume velocity waveform, $g'_{LF}(t)$ (see Figure 5.4(b)). It is a 4-parameter model, that consists of pitch period ($T0$), glottal flow peak position (t_p), instant of maximum closing rate (t_e) and the time constant of an exponential recovery (t_a). The

glottal pulse, $g_{LF}(t)$ is obtained by integrating $g'_{LF}(t)$ with respect to time.

The expressions for the LF-model are shown in Eqn. 5.4 [FLL85]:

$$g'_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_g t), & for\ 0 \le t \le t_e \\ \frac{E_e}{\varepsilon t_a}(e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(T0-t_e)}) & for\ t_e \le t \le T0 \end{cases} \tag{5.4}$$

where E_0 = scale factor, E_e = excitation strength and $\omega_g = \frac{\pi}{t_p}$. The value α can be derived by solving Eqn. 5.5:

$$\int_0^{t_0} g'_{LF}(t)dt = 0 \tag{5.5}$$

Similarly, the value for ε can be derived by solving Eqn. 5.6:

$$\varepsilon t_a = 1 - e^{-\varepsilon(T0-t_e)} \tag{5.6}$$

In terms of discrete sample points, the equations for the LF-model can be re-written in the following form:

$$g_{LF}[n] = \begin{cases} \frac{(e^{\frac{a(n-1)}{N}}[a\sin(\frac{w_a(n-1)}{N}) - w_a\cos(\frac{w_a(n-1)}{N})] + w_a)}{(a^2+w_a^2)}, & for\ 1 \le n \le (N_1+1) \\ e_1[e^{\frac{te-1}{r_b}}(\frac{n-1}{N} - 1 - r_b) + e^{\frac{te-\frac{n-1}{N}}{r_b}}(rb)], & for\ (N_1+1) \le n \le N \end{cases} \tag{5.7}$$

where

$t_e = 0.6$

$N_1 = t_e N$

$w_a = \frac{\pi}{0.7 t_e}$

$a = \frac{-ln[-0.1\sin(w_a t_e)]}{t_e}$

$r_b = 0.1(\frac{1}{a^2+w_a^2}[\frac{1}{0.1}\frac{w_a}{\tan(w_a t_e)} - a + w_a])$

$e_1 = [0.1(1 - e^{\frac{te-1}{r_b}})]^{-1}$

Eqn. 5.7 is adapted from glotlf.m within VOICEBOX, a MATLAB toolbox for speech processing [Bro98]. An example of the glottal waveform generated with the LF-model is shown in Figure 5.4(a). The LF-model has a single discontinuity point in its first derivative waveform (Figure 5.4(b)).

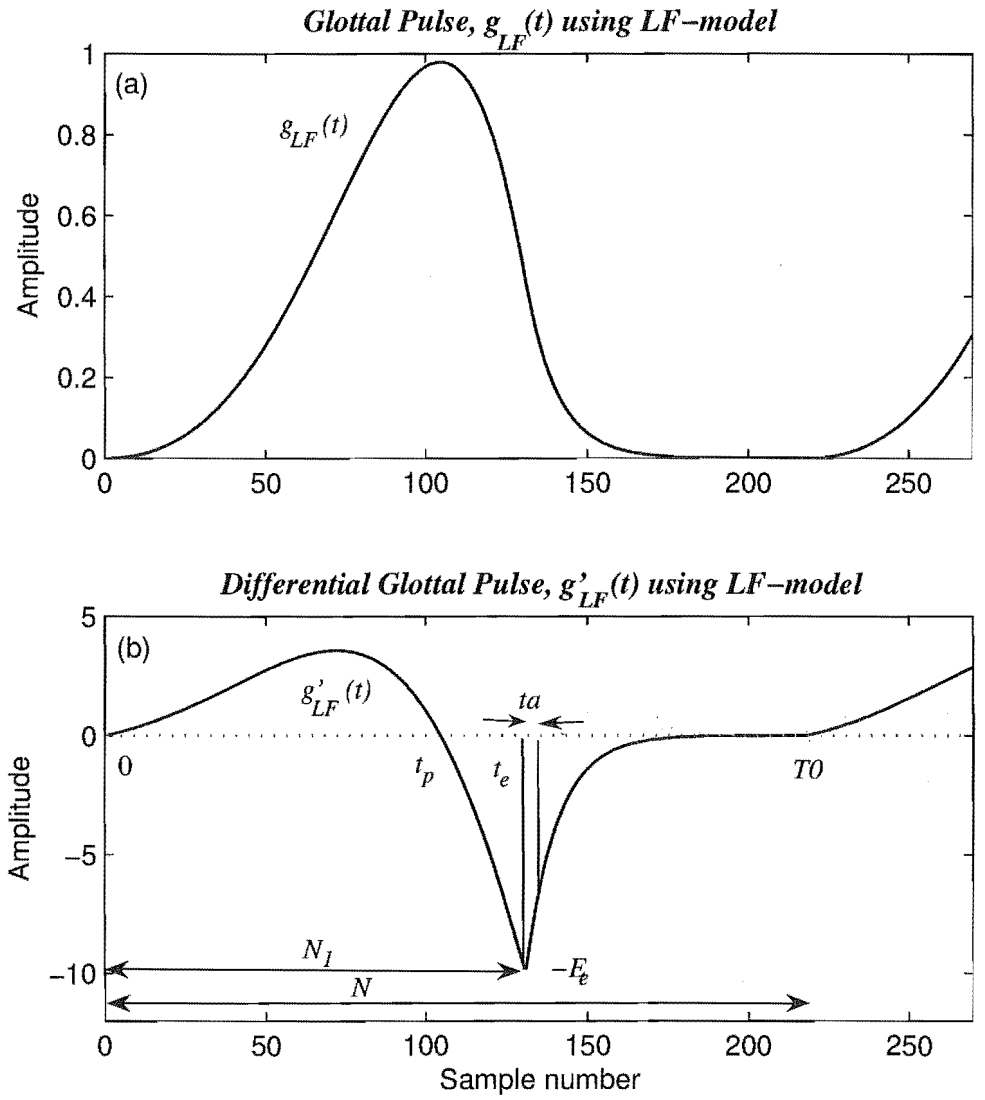

Figure 5.4 The LF-model: (a) glottal pulse, $g_{LF}(t)$ and (b) differential glottal pulse, $g'_{LF}(t)$.

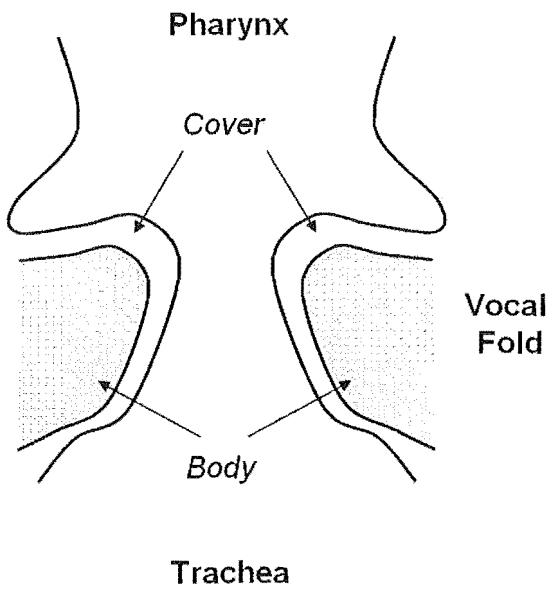

Figure 5.5 Vocal folds, the cover-body structure (based on Hirano [Hir74]).

5.2.4 Physical models

Another approach for producing glottal pulse is to create models that are based on the actual physical properties of the vocal folds vibration. Hirano [Hir74] suggests that the structure of the vocal folds can be divided into two main layers (Figure 5.5): (i) the cover, which consists of a pliable, non-contractile mucosal tissue and (ii) the body layer, that consists of muscle fibers and ligamentous tissue surrounded by the cover layer [Sto02].

Figure 5.6 shows the six stages [(a)-(h)] of a full cycle of the vocal folds vibration in the coronal plane. It shows the motion of the vocal folds are not strictly in the lateral plane but also in the vertical plane. The wave-like motion on the surface of the vocal folds is known as the "mucosal wave".

The vocal folds can be explained as a self-oscillating system. They can convert a steady stream of air from the lungs into a series of flow pulses by the quasi-periodic opening and closing of the glottis due to the Bernoulli effect [dB58].

An early model that based on the physiological structure of the vocal folds is the one-mass glottal model by Flanagan *et al* [FL68, Luc04]. In this model, the mucosal wave and lateral motion of the vocal folds are represented by a single mass-damper-spring system (Figure 5.7(a)). The mathematical representation of the one-mass model is expressed as:

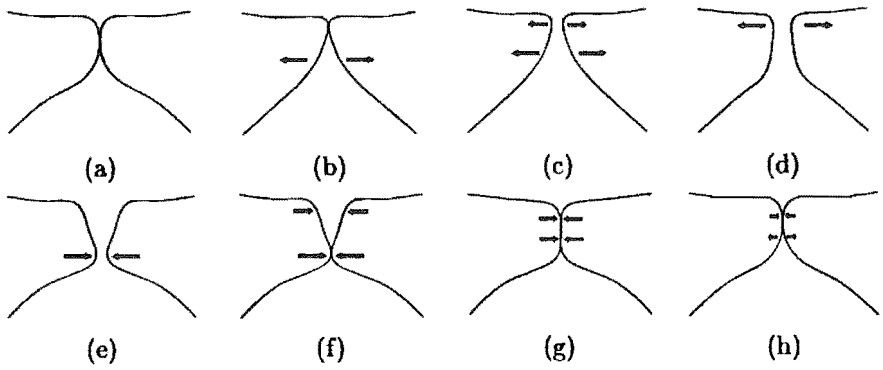

Fig. 4 Diagram showing an idealized cycle of vocal fold vibration in the coronal plane. Note that the lower portion of the vocal folds leads the upper portion creating a wave-like motion on the vocal fold surface. This is called the *mucosal wave*.

Figure 5.6 Mucosal wave. [Sto02]

$$m\ddot{x} + r\dot{x} + kx = dl_g P_g(t) \tag{5.8}$$

where m = mass, r = damping coefficient and k = stiffness coefficient, x = displacement of the mass, d = width of the mass, l_g = length of the mass and $P_g(x)$ = glottal air pressure (glottal pulse).

An improved model from the one-mass model is the two-mass model by Ishizaka *et al* [IF72]. This model takes into account the motion mucosal wave by representing the vocal folds with two mass-damper-spring systems (Figure 5.7(b)). The lower mass (m_1) was made larger and heavier in an effort to simulate the body layer. The expressions for the two-mass model are shown in Eqn. 5.9:

$$\begin{aligned} m_1\ddot{x_1} + r_1\dot{x_1} + k_1x_1 + k_{12}[x_1 - x_2] &= d_1l_gP_{m1}(t) \\ m_2\ddot{x_2} + r_2\dot{x_2} + k_2x_2 - k_{12}[x_1 - x_2] &= d_2l_gP_{m2}(t) \end{aligned} \tag{5.9}$$

where m_n = mass, r_n = damping coefficient and k_n = stiffness coefficient, x_n = displacement of the mass, d_n = width of the mass, l_g = length of the two masses and $P_{m_n}(t)$ = glottal air pressure. The value n = 1 for the bottom mass and 2 for the upper mass. Details of these equations can be found in [KAR99].

Later, a three-mass model by Story *et al* [ST95] was introduced (Figure 5.7(c)). The third mass in this three-mass model was added to represent the body layer. Equations related to this model can be

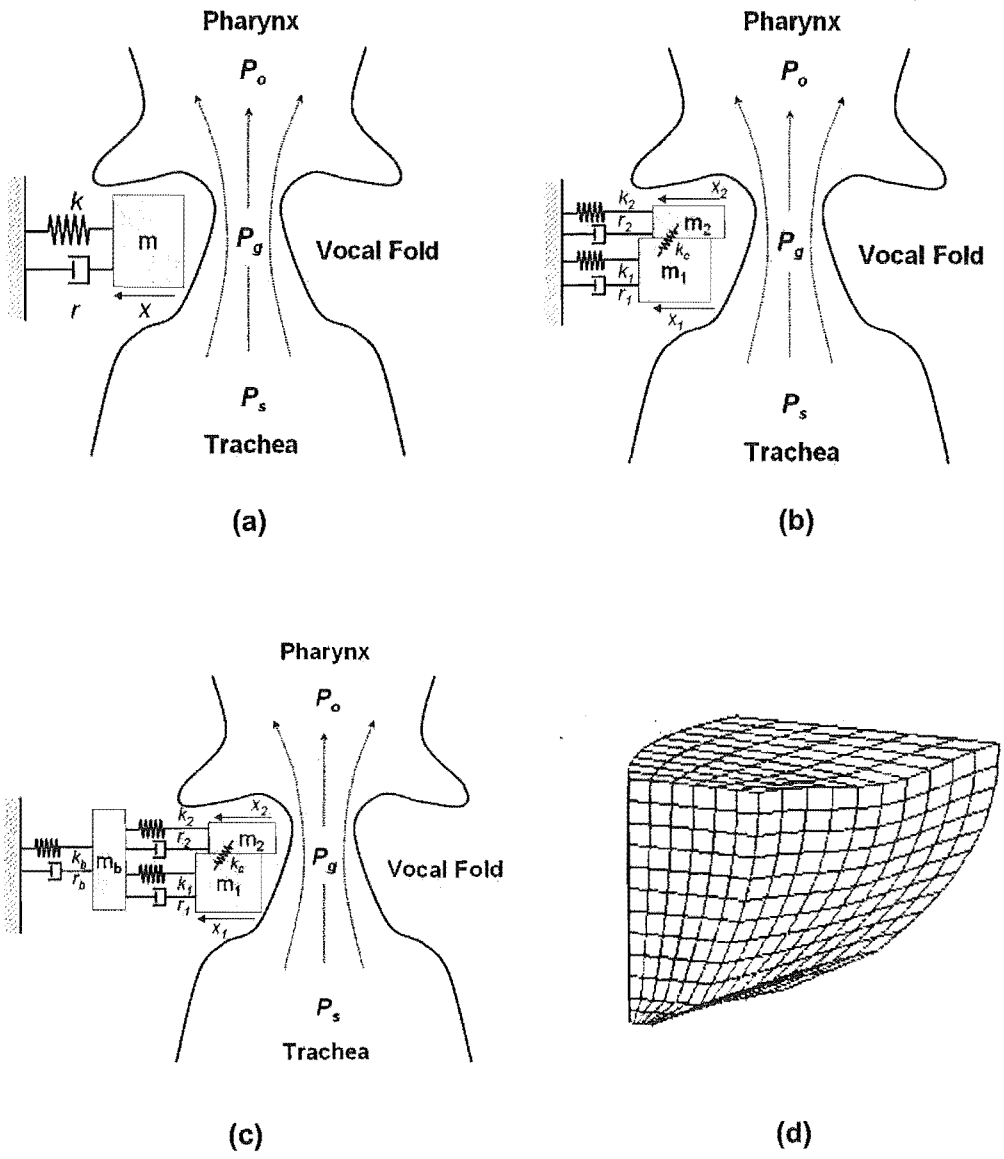

Figure 5.7 (a) the one-mass model [FL68], (b) two-mass model [IF72], (c) three-mass model [ST95] and (d) finite element model [ABT00] (only one 3-dimensional vocal fold is shown here).

obtained from [ST95].

More complex models such as the multi-mass model [KAR99, Tit73, Tit74, TS75] and finite element [ABT00] have also been developed. The finite element model [ABT00], shown in Figure 5.7(d) describes the geometry and physiology of the vocal folds with a large number of small elements. These models are able to provide highly accurate representation of the structure and the mechanics of the vocal folds. But the more accurate a model is, the more complex it is to implement. With the large number of parameters to consider and calculations required, performing real-time simulation with these models is difficult.

Although a physical glottal model has the advantage of generating glottal pulse that is related to the actual physiology of the vocal folds, these models usually requires a large number of parameters to properly characterise the function of the vocal folds. For real-time simulation of the glottal source, such as the sound source for an artificial larynx, it is important to keep the number of parameters to a minimum yet have a model which reliably produces a sufficiently natural sounding voice source.

5.3 The new glottal pulse model: the twin-bar model

The two main problems with the existing glottal models are: (i) even though the glottal waveform shape changes as pitch changes, glottal models such as Rosenberg's model and the LF models retain their shape over all pitch range; (ii) the more complex models such as the physical models produce waveforms that are more realistic but they are not suitable for real-time simulation. Therefore, there is a need for a simple model that is able to change its shape with respect to the change in pitch and is suitable for real-time simulation.

This section introduces a new glottal pulse generator using a conceptual model: the twin-bar model. The twin-bar model differs from other existing models by its ability to track the shape of the glottal waveform (in this case the vocal folds contact area as represented by the EGG signal) as pitch is varied. Most of the commonly used models for speech synthesis (e.g. LF glottal model and Rosenberg's glottal model) only allow a fixed glottal waveform shape, that is, the open quotient (pulse width) and speed quotient (waveform skew) are fixed for all pitch levels. Glottal airflow, *AF*, is the measured airflow from the lungs that is modulated by the opening and closing of the vocal folds. It is sometimes used as glottal sound source for speech synthesis. The new model may be adapted to model glottal airflow by following the same procedures for finding the 3 glottal parameters, namely OP_{AF}, CP_{AF} and sCP_{AF} of a particular waveform. The notation is slightly different to imply that the parameters obtained from airflow may be different from the parameters obtained from EGG parameters.

The *OP*, *CP* and *sCP* parameters used in the twin-bar model were obtained from the experiment that will be discussed in Chapter 7.

5.3.1 Description of the model

A conceptual analogue model has been created to provide a better understanding of the idea behind this new model (refer to Figure 5.8). As the name implies, two rigid bars are the basis of the model.

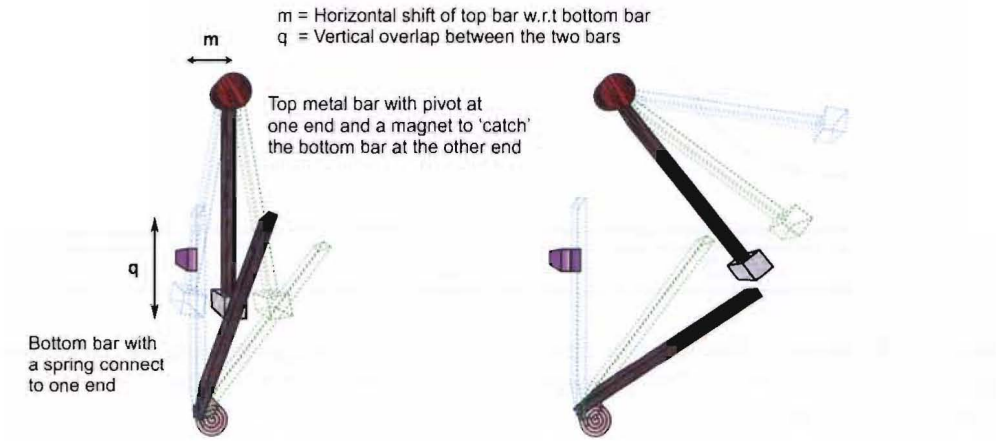

Figure 5.8 The twin-bar model showing the top bar engages on the lower bar causing the lower bar to rotate in the direction of the top bar (left) and the lower bar recoils back to its original position when the two bars disengage (right).

The two bars have unit length. The top bar is pivoted at its upper end at a distance m rightward of the vertical line passing through the pivot of the lower bar. When both bars are vertical, the bars overlap by a constant amount q ($q < 1$). The tip of the bottom bar is defined as the origin, when the bottom bar is in the vertical position. The top bar is driven actively and rotates anti-clockwise. The lower bar is passive and has a spring-loaded pivot at its lower end. A magnet attached to the free end of the top bar engages the lower bar causing it to rotate clockwise until the magnet moves beyond the end of the lower bar. At this point the lower bar recoils back to its original vertical position where it comes to rest against a stop (see Figure 5.8, right).

The waveform generated by the model has 3 consecutive segments. Segment 1 has N_1 samples, segment 2 has N_2 samples and so on. Segment 1 is obtained by measuring the vertical distance between the tip of the bottom bar and the origin, while the two bars are in contact. Figure 5.9 shows a stylised waveform generated by the model. Segment 2 commences when the two bars lose contact. This segment is modelled by a raised cosine function with sCP as its parameter (see Figure 5.9). Segment 3 has the value of zero. Its length is determined by the difference between N (the total number of samples within a cycle) and $(N_1 + N_2)$. The amplitudes of segment 1 and segment 2 were scaled such that the last value within N_1 and the first value of N_2 are the same.

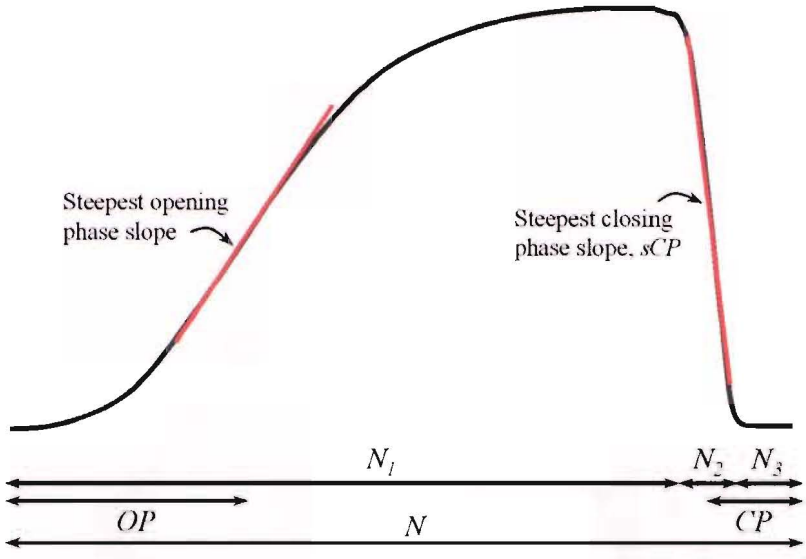

Figure 5.9 Stylised glottal pulse generated by the twin-bar model showing the parameters (*OP*, *CP* and *sCP*), N_1, N_2 and N_3. The glottal pulse was inverted so that it is the same orientation as the EGG signal in Figure 4.12.

5.3.2 Analysis

Although the twin-bar model may not physically model the movement of the vocal folds, it has the ability to produce a range of waveform shapes similar to those observed in recorded EGG signals and glottal airflow signals. It has been reported that waveform shape (e.g. skewness and pulse width) affects the sound that is perceived [Ros71]. The twin-bar model therefore, offers a unique solution for producing waveforms with the desired pulse shape with different sound quality.

Figure 5.10 shows the EGG pulse shapes that can be produced by the twin-bar model (for a fixed *F0*). For the given waveform, the *OP* value can vary between 13 - 136 samples (0.59ms - 6.2ms respectively). Theoretically, m and/or q can be chosen to fit the different glottal waveform shapes. For example, the glottal pulse for falsetto voice is more symmetrical [CL91], it can be simulated by letting m=0 and q=0.8 (see Figure 5.10e). Modal register has a more skewed waveform shape compared with falsetto [CL91], it can therefore be generated by letting m=0.8 and q=0.9. The criteria for choosing the q value for this thesis were based on: (i) the q value that can generate a wide range of *OP* values, and (ii) waveforms generated that visually resemble the glottal waveform of interest (e.g. the EGG waveform). An assumption in the model is that the amplitude of the waveform discontinuity (after scaling) is always equal to 1.

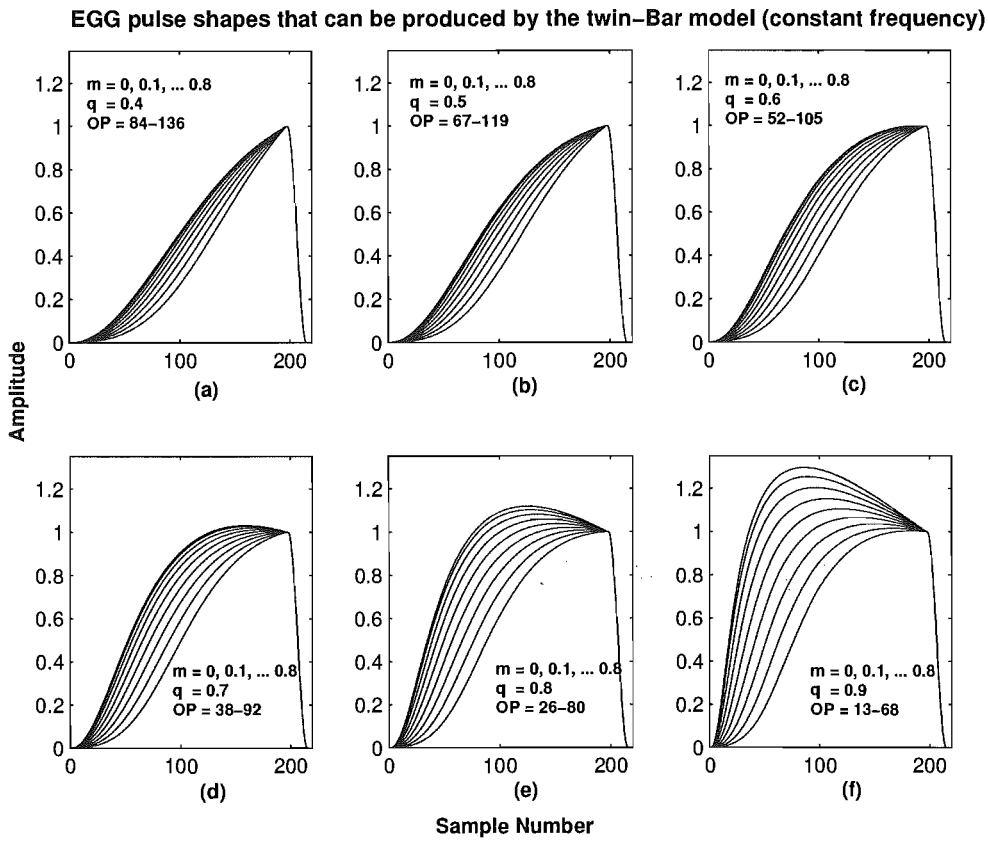

Figure 5.10 A range of EGG pulse shapes that can be produced by the twin-Bar model (constant frequency).

5.3.3 Choice of parameters

The current model uses a sampling frequency, f_s, of 22050 Hz. *F0* is defined as the fundamental frequency in Hz. N is the total number of samples per glottal cycle. The twin-bar model is made up of 3 separate segments, namely N_1, N_2 and N_3, where N is the sum of the three segments. The length of each bar is equal to 1.

A MATLAB program was written to find the q value that will give the biggest *OP* range for $0 \leq m < 1$ numerically. For a fixed waveform length, the biggest *OP* range occurs at q= 0.85 where *OP* varies from around 10% to 50% of the total waveform length of the N_1 segment. The *OP*, *CP* and *sCP* values are shown in Table 5.1. The m equations in Table 5.2 were designed such that they correspond to the *OP* values at a particular *F0* in each group.

Table 5.1 The *OP*, *CP* and *sCP* parameters versus *F0* for male, female and all subjects obtained from experiment.

Group	Male	Female	All
OP	$e^{-1.09lnF0-.70}f_s$	$e^{-1.39lnF0+2.07}f_s$	$e^{-1.24lnF0+.69}f_s$
CP	$e^{-.75lnF0-3.83}f_s$	$e^{-.93lnF0-2.51}f_s$	$e^{-.84lnF0-3.17}f_s$
sCP	$-.08lnF0+.276$	$-.076lnF0+.276$	$-.078lnF0+.276$

Table 5.2 The m versus *F0* equations obtained from the *OP* equation.

Group	m
Male	$-1.0x10^{-8}F_0^3+1.4x10^{-5}F_0^2-.0081xF_0+2.1$
Female	$-9.6x10^{-8}F_0^3+5.8x10^{-5}F_0^2-.014xF_0+1.7$
All	$-3.0x10^{-8}F_0^3+2.8x10^{-5}F_0^2-.011xF_0+1.9$

The first segment of the twin-bar model, N_1 is as follows:

$$N_1 = [N - CP - \frac{\pi}{2\Delta_2}]_{round} \tag{5.10}$$

$$\Delta_2 = -\sin^{-1}(2sCP) \tag{5.11}$$

where N = length of the cycle. The value Δ_2 is the step size (in terms of sample points) for the segment defined by the second segment, N_2. *CP* and *sCP* were obtained from the equations in Table 5.1, depending on gender.

The second segment of the twin-bar model, N_2 is defined as:

$$N_2 = [N - CP + \frac{\pi}{2\Delta_2}]_{round} - N_1 \tag{5.12}$$

The the minimum angle (θ_{min}) and maximum angle (θ_{max}) of the top bar are as follows:

$$\theta_{min} = \sin^{-1}(-m) \tag{5.13}$$

$$\theta_{max} = \cos^{-1}(\frac{x}{2}) - \sin^{-1}(\frac{m}{x}) \tag{5.14}$$

where $x = \sqrt{m^2 + (2-q)^2}$ and $q = 0.85$.

The step size Δ_1 (in terms of number of sample points) for segment 1, is:

$$\Delta_1 = \frac{\theta_{max} - \theta_{min}}{N_1} \tag{5.15}$$

The scale factor, k was chosen to allow for continuity of the $g_{TB}[n]$ signal:

$$k = -\cos(\tan^{-1}[\frac{\sin(\theta_{max}) + m}{2 - q - \cos(\theta_{max})}]) + 1 \tag{5.16}$$

The exponential equations for *OP* versus *F0* obtained from experiment were used to find the m versus *F0* equations (via regression) for each group listed in Table 5.2. Since the bars on the twin-bar model are set to the length of 1, m values are limited to between 0 and 1. This has the effect of limiting the *F0* range that can be generated by each group (see Table 5.3).

The expressions for the twin-bar model are shown in Eqn. 5.17:

Table 5.3 Pitch range that the twin-bar model can cope with for different groups

Group	*F0* range (Hz)
Male	65-270
Female	195-590
All	120-410

$$g_{TB}[n] = \begin{cases} \frac{1}{k}[1-\cos(\tan^{-1}[\frac{\sin(n\Delta_1+\theta_{min})+m}{2-q-\cos(n\Delta_1+\theta_{min})}])], & 0 < n \leq N_1 \\ \frac{1}{2}[1+\cos([n-(N_1+1)]\Delta_2)], & N_1 < n \leq N_1+N_2 \\ 0, & N_1+N_2 < n \leq N \end{cases} \quad (5.17)$$

5.3.4 Comparison to earlier models

The twin-bar model is able to produce pulse shapes with the *OP* duration longer than the *CP* duration, an important characteristic of glottal pulse. The pulse shape produced by the two-pole model always has the *OP* duration shorter than the *CP* duration. Unlike the Rosenberg and LF models, the twin-bar model can produce a pulse shape that varies with pitch. Compared with physical models, the twin-bar model is simple as it only requires a single input parameter: pitch, and can be generated in realtime.

5.4 Vocal tract modelling

This section covers the vocal tract model, the structure used for the implementation of the model and the digital filter realisation.

5.4.1 Acoustic tube model

The vocal tract is part of the speech production mechanism. The shape of the tract alters the sounds that are uttered. The resonances within the tract give rise to the formants that correspond to the different vowel sounds. The vocal tract can be modelled as a series of concatenated lossless cylindrical tubes with equal length, L, and with varying cross-sectional area, A_i, where $i = 1,2,3,...N$ (see Figure 5.11). Sodhi [Son86] found that the curvature of a uniform tube changes the points of resonance by only a few percent from those of a straight tube. Experiments by Schroeter *et al* [SS94] have also shown that the curvature of the vocal tract does not affect the sound propagation appreciably.

Digital waveguide modelling is a computational modelling technique commonly used in speech synthesis, acoustics and computer music. This technique assumes that the sound wave travels as a one-dimensional plane. The model for plane wave propagation in the vocal tract is valid for frequencies up to 3.5kHz [SS94]. Above this frequency, the vocal tract diameter is no longer much smaller than the wavelength and therefore transverse modes can propagate. Speech energy is almost totally concentrated below 3.5kHz.

The one-dimensional wave equation in a lossless cylindrical tube is given by:

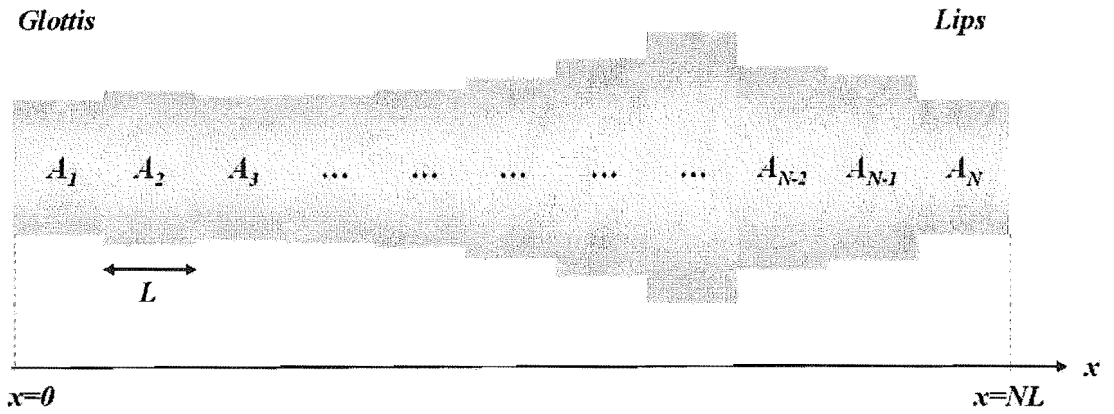

Figure 5.11 The concatenated lossless cylindrical tube vocal tract model.

$$\frac{\partial^2 p}{\partial t^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial x^2} \tag{5.18}$$

where p = pressure displacement in the tube, t = time, c = speed of sound ($\approx 345ms^{-1}$) and x = distance along the tube axis. The derivation of Eqn. 5.18 is obtained from Newton's second law of motion. Details of this derivation can be found in [FR91].

The D'Alembert's general solution to the one-dimensional wave equation [Ava02, FR91] is of the form:

$$p(x,t) = p^+(x-ct) + p^-(x+ct) \tag{5.19}$$

where $p^+(x-ct)$ is the forward propagating pressure wave (moving to the right) with velocity c, and $p^-(x+ct)$ is the backward propagating pressure wave (moving to the left) with velocity $-c$.

The characteristic impedance of an acoustic tube, Z_0, is defined by [JHP00]:

$$Z_0 = \frac{\rho c}{A} \tag{5.20}$$

where ρ = density of air in vocal tract ($1.14kgm^{-2}$) and A = the cross-sectional area of the tube. In terms of pressure and volume velocity, Z_0 is defined as [FR91]:

$$Z_0 = \frac{p^+}{u^+} = \frac{p^-}{u^-} \tag{5.21}$$

where u = particle velocity.

5.4.2 The Kelly-Lochbaum structure

When there is a discontinuity between one tube and the next, the characteristic impedance changes at the interface. A forward propagating waveform will be partially reflected back into the tube, a phenomenon known as *scattering* [Väl95]. Scattering only occurs at the interfaces.

Assuming continuity of pressure, p, and conservation of volume velocity, u, at the interface between the k^{th} and $(k+1)^{th}$ tube:

$$p_k(L,t) = p_{k+1}(0,t) \tag{5.22}$$

$$u_k(L,t) = u_{k+1}(0,t) \tag{5.23}$$

The total sound pressure in the k^{th} tube section is expressed as:

$$p_k(x,t) = p_k^+(x-ct) + p_k^-(x+ct) \tag{5.24}$$

where x lies within the k^{th} segment. The total volume velocity in the k^{th} segment is the difference between the two components divided by Z_k, the characteristic impedance of the k^{th} tube:

$$u_k(x,t) = \frac{1}{Z_k}[p_k^+(x-ct) - p_k^-(x+ct)] \tag{5.25}$$

again for x lying within the k^{th} segment. Introducing $\tau = \frac{L}{c}$, the time for a wave to traverse a section, by substituting Eqn. 5.24 into Eqn. 5.22, the following equation is obtained:

$$p_k^+(t-\tau) + p_k^-(t+\tau) = p_{k+1}^+(t-\tau) + p_{k+1}^-(t+\tau) \tag{5.26}$$

Similarly, substituting Eqn. 5.25 into Eqn. 5.23 with $\tau = \frac{L}{c}$ yields:

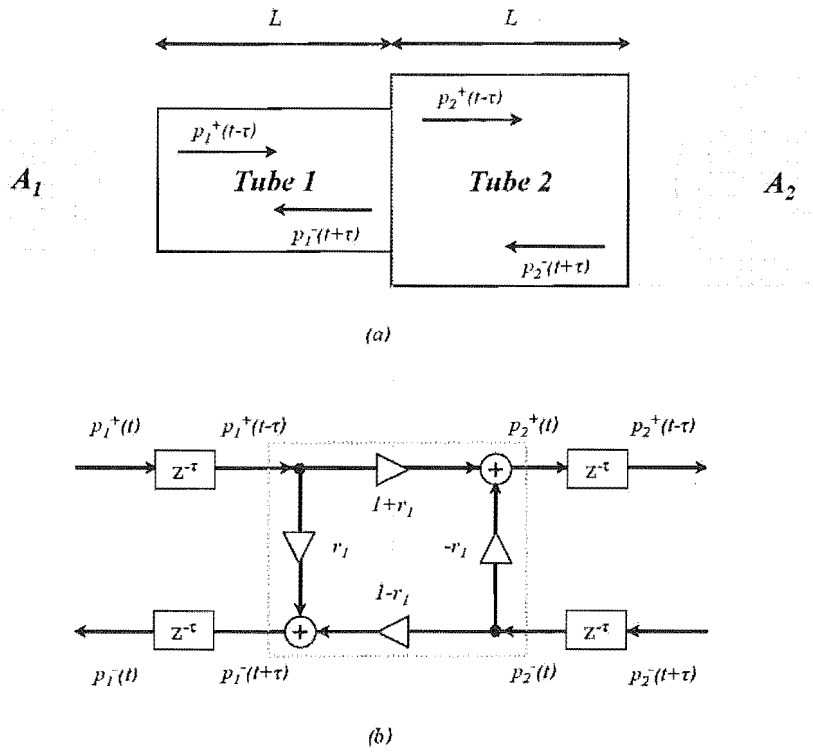

Figure 5.12 The Kelly-Lochbaum scattering junction [KL62].

$$\frac{1}{Z_k}[p_k^+(t-\tau)-p_k^-(t+\tau)]=\frac{1}{Z_{k+1}}[p_{k+1}^+(t)-p_{k+1}^-(t)] \tag{5.27}$$

Solving for $p_k^-(t+\tau)$ and $p_{k+1}^+(t)$, the scattering equations become:

$$p_k^-(t+\tau)=r_k p_k^+(t-\tau)+(1-r_k)p_{k+1}^-(t) \tag{5.28}$$

$$p_{k+1}^+(t)=(1+r_k)p_k^+(t-\tau)-r_k p_{k+1}^-(t) \tag{5.29}$$

where r_k is the reflection coefficient at the interface between the k^{th} and $(k+1)^{th}$ tube given by:

$$r_k=\frac{Z_{k+1}-Z_k}{Z_{k+1}+Z_k}=\frac{A_k-A_{k+1}}{A_k+A_{k+1}} \tag{5.30}$$

The reflective coefficients for the two ends of the vocal tract are defined as r_g (reflective coefficient at the glottis) and r_l (reflective coefficient at the lips). These values were defined by performing a perceptual test of the voice generated using the glottal models (Rosenberg's model, LF model and the twin-bar model) and the vocal tract model. The values r_g = 0.8 and r_l = -0.8 were determined to result in the best voice quality by the author, for all three models.

The scattering equations (Eqns. 5.29 and 5.28) were first derived by Kelly *et al* [KL62] for the design of an acoustic tube model for speech synthesis. The Kelly-Locbaum scattering junction is shown in Figure 5.12. For a 3-segment vocal tract model, the scattering junctions are shown in Figure 5.13.

A more efficient implementation of the scattering junction is to rearrange Eqns. 5.29 and 5.28 to form an equivalent one-multiply scattering junction [MG76], expressed as:

$$p_k^-(t+\tau) = p_{k+1}^-(t) + r_k[p_k^+(t-\tau) - p_{k+1}^-(t)] \tag{5.31}$$

$$p_{k+1}^+(t) = p_k^+(t-\tau) + r_k[p_k^+(t-\tau) - p_{k+1}^-(t)] \tag{5.32}$$

Since the term $r_k[p_k^+(t-\tau) - p_{k+1}^-(t)]$ is the same for both Eqns. 5.31 and 5.32, it has to be computed only once. The configuration of the one-multiply scattering junction with its single multiplier is shown in Figure 5.14. The multi-tube vocal tract model can now be implemented with only digital delay lines and one-multiply scattering junctions. For digital computation, the length of each tube, L, is chosen such that $L = cT0$, with τ equals to $T0$, the sampling interval. Then $z^{-\tau}$ becomes a unit delay.

The advantage of modelling the vocal tract with digital waveguide technique is that this model behaves like a physical system. It is possible to work out the waveform propagation at a given time and space. Since the model is assumed to be linear, it can be realised in either the frequency or the time domain.

5.4.3 Half-sample delay Kelly-Lochbaum structure

In the half-sample Kelly-Lochbaum delay model, each segment corresponds to a delay of $z^{-\tau}$ where $\tau = \frac{T0}{2} = (\frac{1}{2f_s})$ instead of $T0$. The length of each segment, L, is calculated with the following equation:

$$L = \frac{c}{2f_s} \tag{5.33}$$

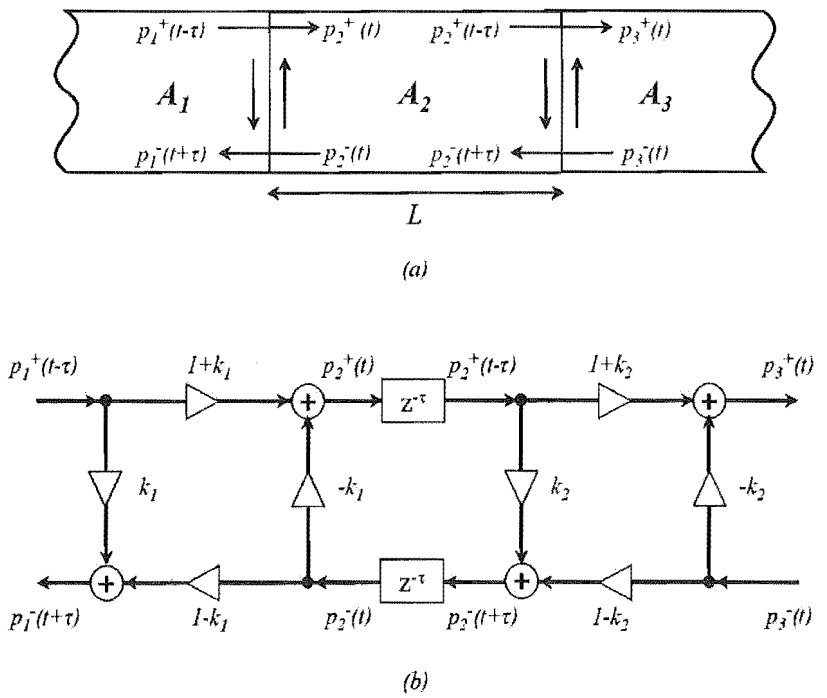

Figure 5.13 The waveguide digital structure with scattering junctions for a 3-tube model.

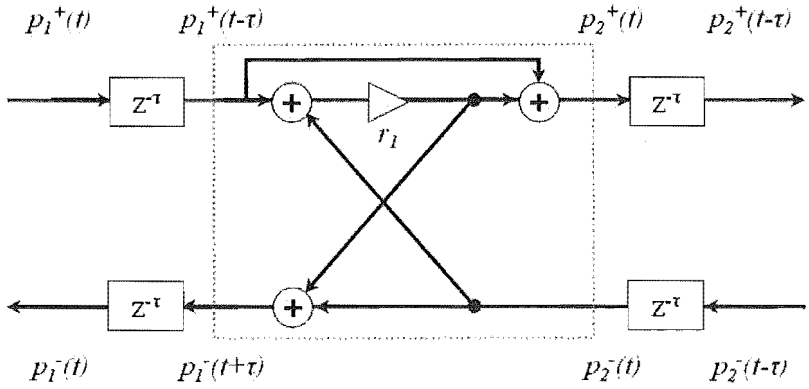

Figure 5.14 The one-multiply scattering junction [MG76].

where c = speed of sound in the vocal tract and f_s = sampling rate.

Given a fixed f_s, the half-sample delay model allows the vocal tract to be modelled with twice the number of segments (e.g. twice the spatial resolution) compared with the unit sample delay version. This is important for speech synthesis where an accurate vocal tract model is desired. The downside of the half-sample delay model, however, is it comes with twice the computational cost.

5.4.4 Z-transform of the lossless tube model

The lossless tube model can also be implemented in the frequency domain. Figure 5.15(a) shows the signal flow diagram of the lossless tube model's equivalent discrete-time system. The discrete-time signal flow model in Figure 5.15(a) can be modified to have a whole-delay sample instead of half-delay sample. Moving $z^{\frac{-N}{2}}$ delay (where N is the number of tubes) into the forward path introduces an additional delay of $\frac{N}{2}$ samples. The modified flow diagram is equivalent to the flow diagram in Figure 5.15(a) provided that the signal is forward by $\frac{+N}{2}$ samples after the lip radiation as shown in Figure 5.15(b).

The transfer function of an N-tube lossless vocal tract model can be expressed as [IES05]:

$$H(z) = \frac{G}{D(z)} = \frac{1+r_g}{2} \cdot \frac{\prod_{k=1}^{N}(1+r_k)z^{\frac{-N}{2}}}{D(z)} \tag{5.34}$$

where

$$D(z) = \begin{bmatrix} 1 & -r_g \end{bmatrix} \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -r_N \\ -r_N z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{5.35}$$

If $r_G = 1$, $D(z)$ can be computed recursively with the following equations:

$$D_0(z) = 1$$
$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}),\ k = 1, 2, \ldots, N$$
$$D(z) = D_N(z)$$

The reflection coefficient at the tube junction between tube *k+1* and tube *k* is given by Eqn.5.30 and is repeated here for convenience:

$$r_k = \frac{A_k - A_{k+1}}{A_k + A_{k+1}} \tag{5.36}$$

The reflective coefficient at the lips is defined as:

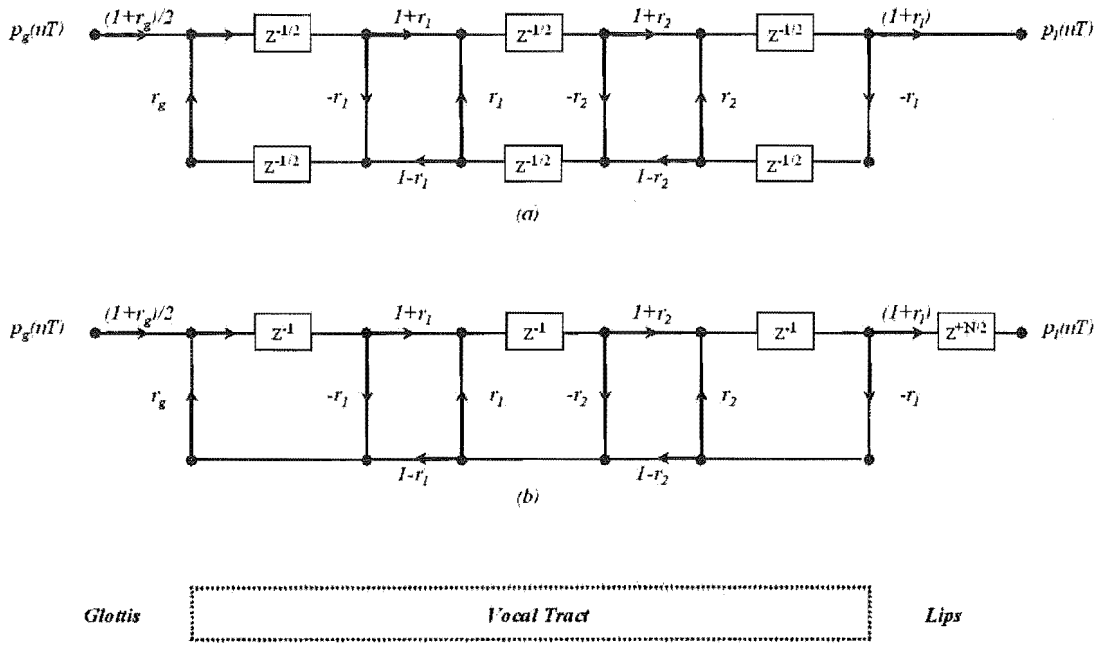

Figure 5.15 (a). The signal flow diagram of the lossless tube model's equivalent discrete-time system (for N = 4). (b). The equivalent discrete-time system using only whole delays in the ladder part.

$$r_l = r_N = \frac{A_N - A_{N+1}}{A_N + A_{N+1}} \tag{5.37}$$

The vocal tract transfer function (Eqn. 5.34) can be simplified to an all-pole filter design:

$$H(z) = \frac{H_0}{1 - \sum_{k=1}^{N} b_k z^{-k}} = \frac{H_0}{\prod_{k=1}^{N} (1 - p_k z^{-1})} \tag{5.38}$$

where H_0 is the overall gain term and p_k is the complex pole locations of the N-tube model in the the Z-plane. Each pair of complex conjugate poles represent one of the formants of the vocal tract spectrum. Provided that the poles are well separated, the formant frequency and bandwidth are approximated by:

$$\mathbf{F}_k = \left(\frac{f_s}{2\pi}\right) \tan^{-1}\left[\frac{Im(p_k)}{Re(p_k)}\right] \tag{5.39}$$

$$\mathbf{B}_k = -\left(\frac{\mathscr{F}_k}{\pi}\right) \ln|p_k| \tag{5.40}$$

where $\mathbf{F}_k$ and $\mathbf{B}_k$ are the k^{th} formant and bandwidth respectively.

5.4.5 Lip radiation

The radiation of sound at the lips is represented by the transfer function, *R(z)*. Flanagan [Fla72] showed an accurate portrayal of the impedance at the lips by representing it as:

$$R(\omega) = \frac{k_1 k_2 \omega}{\omega^2 k_2^2 + k_1^2}(\omega k_2 - j k_1) \tag{5.41}$$

where $k_1 = \frac{128}{9\pi^2}$, $k_2 = \frac{8r}{3\pi c}$ and r is the radius of the opening at the lips.

The magnitude and phase of the lip impedance are defined as:

$$|R(\omega)| = \frac{\omega k_1 k_2}{\sqrt{k_1^2 + \omega^2 k_2^2}} \tag{5.42}$$

$$Phase\{R(\omega)\} = \tan^{-1}(\frac{k_1}{\omega k_2}) \tag{5.43}$$

At $\omega = 0$, $R(0) = 0$. As frequency increases, the magnitude of $R(\omega)$ increases. Clearly then, $R(\omega)$ has a high-pass filtering effect, so for simplicity it can be represented with a simple differentiator [JHP00] as shown in Eqn. 5.44.

$$R(z) = 1 - z_0 z^{-1} \tag{5.44}$$

where z_0 is a constant less than 1 to ensure stability when inverse filtering is performed. For voice synthesis in this thesis, $z_0 = 0.99$ was used.

5.4.6 Choice of vocal tract model for voice synthesis

Both the half-sample Kelly-Lochbaum and the simplified Z-transform options have been used for preliminary voice synthesis. The half-sample delay Kelly-Lochbaum structure was found to be better for the following reasons:

(a). It is easier to understand the wave propagation process in the system.
(b). The half-sample delay Kelly-Lochbaum technique has better spatial and time resolution.

For the above reasons, the time domain voice synthesis technique will be used in the remaining chapters of this thesis. The algorithm for the half-sample delay Kelly-Lochbaum approach is detailed in Chapter 7.

5.5 Summary

In this chapter, a review of well-known glottal pulse models was presented. A new glottal pulse model, the twin-bar model was also introduced. The twin-bar, Rosenberg and LF models will be used for comparing the quality of the synthesised voice generated with the twin-bar model in Chapter 7.

Two vocal tract models based on waveguide model were also discussed. They provide a consistent vocal tract shape (filter) for the glottal pulse model in voice synthesis. The vocal tract model that will be used for the remaining chapters of this thesis is the half-sample delay Kelly-Lochbaum approach (Chapter 7).

Chapter 6

Alternative pitch control methods for the voiceless

Pitch, voice intensity and voice type are the three key factors that affect the naturalness of voice. The voice generated by an artificial larynx is usually monotonous because the voice source is simulated using periodic signals at a fixed fundamental frequency (*F0*). The *F0* of a periodic signal is defined as the inverse of the time taken to complete a cycle of the signal. The simulated voice sounds unnatural because normal vocal fold vibration, as shown in the voice literature, is quasi-periodic with a small degree of cycle-to-cycle frequency and amplitude variations. In addition to these short-term variations, the human voice allows for a pitch range of up to 3 octaves [CC96, HJ73], with normal speaking range concentrated in the lower part of a speaker's total range [Fry79]. Baken and Orlikoff [BO00] quote the results of a number of previous studies where the spontaneous speech *F0* in normal male subjects was reported to range from 116 to 123.3Hz with SD of 2.64 to 3.4 semitones. Other studies by Traunmüller *et al* [TE93] and John-Lewis [JL86] reported similar values for conversational mode. In running speech, *F0* varies with a standard deviation estimated to be approximately 3.4 semitones for males and 2.7 semitones for females [JL86]. It is likely, therefore, that long-term *F0* variations may be required for voice simulation, in addition to cycle-to-cycle *F0* perturbation, to attain a natural speech quality in voice synthesis.

It is noteworthy that different vowels, even when produced in isolation, have been found to differ in *F0*, known as intrinsic *F0* [Ewa79, LJNH00, Möb03, Pet78, WL95]. A study of vowels in 31 languages concluded that intrinsic *F0* is not from a deliberate enhancement of the signal but a direct result of vowel articulation [WL95]. High vowels such as /i/ and /u/ have higher *F0* than low vowels like /a/ [TE93, WGKH98, WGL99, Zee80]. Whalen *et al* [WL95] reported that the overall mean vowel *F0* for the 31 languages studied was 177.4 Hz for /u/, 174.9 Hz for /i/ and 160.9 Hz for /a/, with the difference between the high vowel and low vowel to be 15.3 Hz across the languages. Petersen [Pet78] reported a 10-35 Hz range for intrinsic *F0*. As English is a non-tonal language,

we would expect the long-term *F0* variation due to (non-emphasised) running speech to be largely dependent on the vowel identity.

In order to attain a natural speech quality in voice synthesis, long-term *F0* variations are needed, in addition to the cycle-to-cycle *F0* perturbation. While random variation is applicable in simulating short-term *F0* variation, the relatively large long-term *F0* variation in natural speech may be better simulated through a continuous monitoring scheme. This chapter explores the different ideas of pitch control on normal subjects. The most reliable method is then used as feedback for simulation of time-varying *F0* variation to achieve a higher degree of naturalness in artificial voice.

6.1 Pitch control methods considered

Several pitch control techniques were considered including the measurement of thumb pressure, eyebrow movement, laryngeal muscle movement, laryngeal height and jaw movement. To test the feasibility of these ideas, some of the designs were built and tested.

6.1.1 Thumb pressure sensor

This method measures the pressure exerted by the thumb on a pressure-sensitive pad. The thumb pressure is translated into an electrical signal which is used to control the change in *F0* of the artificial voice source. While this pitch control technique may be suitable for some patients, ICU patients who are on ventilator are usually too weak to hold the device. Controlling a pressure-sensitive device requires a fairly precise hand control. It requires too much effort from these patients and is not very practical since the user has to make a conscious effort to press the sensor in order to change the pitch.

6.1.2 Eyebrow movement

Facial muscle functions for most patients are usually unaffected by their physical state. This suggests that a voluntary movement of the facial muscles, such as eyebrow movement, may be an option for pitch control. Patients could have electrodes placed on the eyebrow muscles to measure the eyebrow movement as they speak. The eyebrow movement measurement would then be used to control the pitch of the artificial larynx.

This idea may work in theory but in practice the muscle movement is too coarse for continuous pitch variation. The fact that eyebrow movement is a voluntary movement means that patients who use this device will not only have to think of what they want to say but also how to control the eyebrow movement. In the end, patients will most likely abandon the pitch control and continue speaking without pitch variation. Furthermore, patients who use the device may feel self-conscious about the way they look when they have to keep moving their eyebrows as they speak. Taking all these facts

into consideration it is clear that the eyebrow movement is more suitable to be used to control a switch. For example, it could be used to turn the artificial larynx on/off when required.

6.1.3 Laryngeal muscle movement

A number of investigators [Arn61, FA57, HOV69] reported that an increase in electromyograph (EMG) activity of the cricothyroid muscle (CT) accompanied the increase in *F0* in human. This suggests that it may be possible for ventilator dependent patients to use their CT muscle to control the pitch of an artificial larynx, provided that the mapping between muscle activity and pitch level is available.

This is, however, an invasive option as it requires the insertion of EMG probes into the muscles (located deep inside the neck) to measure the electrical signals sent from the brain to the muscles. Not many people will be keen on this option if there are other non-invasive options available. Besides, the laryngeal muscle technique will not be suitable for laryngectomees who have had their larynx, including the cricothyroid and its surrounding muscles, removed.

Although there is a less invasive EMG measurement technique which uses surface mount electrodes, they also are unsuitable because the muscles that control pitch (intrinsic muscles) are located within the thyroid cartilage. Surface electrodes will only be able to measure the activities of strap muscles (extrinsic muscles) which control voice tension, not related to pitch change.

6.1.4 Laryngeal height

Research has shown that when a person changes their pitch, their laryngeal height alters [SH72]. A prototype of a laryngeal height sensor was built by the author and tested. Is consists of a brass lever structure where the 'arm' of the lever is placed below the larynx. A magnet is placed on the lever 'arm' and a linear hall-effect sensor is mounted on a stable surface below the lever. The linear Hall effect sensor measures the magnetic field as the lever 'arm' moves when the subject speaks. Figure 6.1 shows the prototype laryngeal height sensor.

The sensor was able to act as a pitch control mechanism when tested on normal subjects. In this case, lower frequency corresponds to lower laryngeal height and vice versa. This device, however, fails to work on tracheostomised patients because the breathing tube placed just below the larynx prevents the larynx from moving.

6.1.5 Random pitch variation

Random pitch variation or repeating a pitch profile using pre-stored pitch variation is the easiest pitch control method. This method is often used in text-to-speech software where the artificial speech seems to have a repeated rhythm. However, because the pitch variation is uncorrelated to the

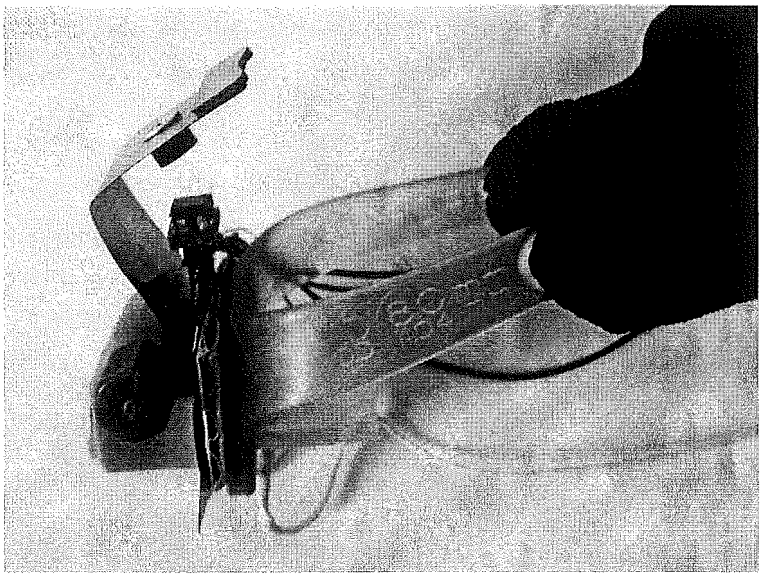

Figure 6.1 The prototype laryngeal height sensor built by the author.

words that is produced, the resulting speech sounds unnatural.

6.1.6 Jaw movement

The main problem with the methods discussed above (except for the laryngeal height option) is the lack of natural control. The users have to make a conscious and unnatural effort to vary the pitch as they speak.

One way around the problem is to consider an interesting option where part(s) of the body that naturally move as an individual speaks are monitored and used as feedback for pitch control. These body parts are mainly located at the head region, such as the lips, tongue and jaw. Pitch control by means of jaw movement was considered because jaw movement measurement is straightforward, cheap and non-invasive. The next section discuss in detail a preliminary study that was carried out to determine the suitability of jaw movement as a way of producing natural pitch control.

6.2 Preliminary study on vocal folds and jaw movement in voiced sounds

The relationship between the magnitude of jaw opening and the vocal parameters in normal speech has not been well studied, although there has been indication through cineradiographic observation that *F0* is related to the position of the mandible [ZG89]. The study of jaw opening during speech is often concerned with the effects of emotion and based on a single vowel. Recent studies

on the articulatory characteristics of speech (particularly mandible) and prosody have shown that jaw opening increases with increased irritation, anger, emphasis or vocal intensity [Möb03, EH96, EFP98, EMHD00, Geu01, WF89]. To minimise extraneous factors related to sporadic or drastic voice changes, we focused on the observation of vowel-related vocal behaviours using vowels embedded in non-emphasised speech with neutral emotion. It has also been observed with high-speed imaging of vocal fold vibrations that different vowels are associated with different vocal fold vibration pattern and larynx position [MHG96]. A study of German speakers showed that *F0* tends to increase from opened to closed vowels and that vowels differ on measures of open quotient (*OQ*), a ratio of open phase to cycle length, and speed quotient (*SQ*), a ratio of opening phase and closing phase [Mar96]. Based on these preliminary findings of vowel-related variation in jaw and vocal fold vibration pattern, we proposed to examine the relationship between vowel identity and the magnitude of jaw opening and glottal parameters.

To start exploring the usefulness of a jaw-tracking device in facilitating better voice simulation in an artificial larynx, this study investigated the possible relationship between voice source parameters and the magnitude of jaw opening associated with vowel change. The voice source, for the purpose of this study, is characterised by the average *F0*, *OQ*, and *SQ* derived from the electroglottographic (EGG) signal. Henrich *et al* [HRC03] used this same technique to find the *OQ* values for their experiment. It is not our intention to associate *OQ* and *SQ* measures obtained from the EGG signals with *OQ* and *SQ* parameters obtained from airflow signals, but to use these measures as indicators of how vocal fold contact patterns vary with different vowels. The measures may also be useful for modelling EGG signals and in the instance where the EGG signal is used to form a glottal excitation waveform for voice synthesis.

Jaw movement in the sagittal plane has been shown to involve a combination of rotational and translational movement [DH02] with the range of (vertical) jaw opening between 6.84 - 11.20mm [EFP98]. Although multiple sensors can be employed to measure both rotational and translational movement, we sought to use a straightforward single sensor, which suggested measuring either the vertical or the horizontal jaw translational movement. A study by Erickson *et al.* [EFP98] measured jaw movement by taking the lowest vertical position of the jaw (maximum jaw opening) at a particular vowel sound, presumably because vertical jaw movement is more evident than horizontal jaw movement. In our experiment, we also measured the vertical jaw movement, since that appeared to be the single component most sensitive to the vowels being uttered.

6.2.1 Method

Subjects and Subjects' Tasks

The subjects were ten New Zealanders of European descent with English as their native language and with no history of speech, voice, or hearing problems. These subjects were recruited through an

advertisement in a university campus. As this was a preliminary study of the relationship between jaw opening and *F0* variations, only male volunteers were included. The age of the subjects ranged from 19 to 58 years (Mean = 31.9 years, SD = 15.7).

Subjects were asked to read aloud, for each experimental trial, a sentence embedded with a meaningless word composed of an isolated vowel (V), a consonant-vowel syllable (CV), or a vowel-consonant syllable (VC). The vowels used were /a/, /e/, /i/, /o/, and /u/. The consonants used were voiced and voiceless plosives, including /b/, /d/, /g/, /p/, /t/, and /k/. The sentence was 'say _ again' (e.g., 'say a again', 'say boo again'). The use of a sentence was to avoid different use of intentional stress on the words presented to the subjects as it has been reported that jaw opening increases with emphasised syllables [WF89]. The subject was asked to speak in a normal conversational manner during the experiment.

6.2.2 Instrumentation

A multi-channel digital recording system was configured to simultaneously record three signals: acoustic signal, jaw position signal, and EGG signal (Figure 6.2). A mounting bracket attached to an adjustable headband was used to attach transducers and is henceforth referred to as "the head mount". The head mount was important because we needed to limit the subject's head movement to avoid any spurious recording during the experiment.

For acoustic recording, a condenser microphone (AKG C420, AKG Acoustic GmbH, Austria) fixed on the front of the head mount was connected to a microphone amplifier (Eurorack MX602A from Behringer, Germany). For the recording of jaw opening, a custom-made unit, which used a spring-loaded potentiometer with a linear working range of 11mm, was used (Figure 6.2). The potentiometer was connected in a voltage divider configuration, with a supply voltage of 5V dc. For EGG recording, a commercial electroglottograph (Kay Elemetrics Model 6103, USA) was used.

The outputs of the microphone amplifier, the jaw potentiometer, and the electroglottograph were connected to three separate channels of an A/D converter (DAQCard-AI-16E-4, National Instruments, USA) via a SCB-68 68-pin shielded connector box. The A/D converter was interfaced to a laptop computer (Compaq (Taiwan) 650MHz Pentium 4). The MATLAB (The Mathworks, Inc.) Data Acquisition Toolbox was used to acquire the digitised data from the A/D converter at 4000 samples per second. The signals were acquired at 12-bit resolution but were stored as .wav files with 8-bit resolution for compatibility with the software used. Programs developed by the author written in MATLAB12 were installed for signal acquisition and analysis. The relatively low sampling rate for the acoustic signal was sufficient since this signal was only used to verify acoustically when each vowel was being uttered instead of being used for actual waveform analysis.

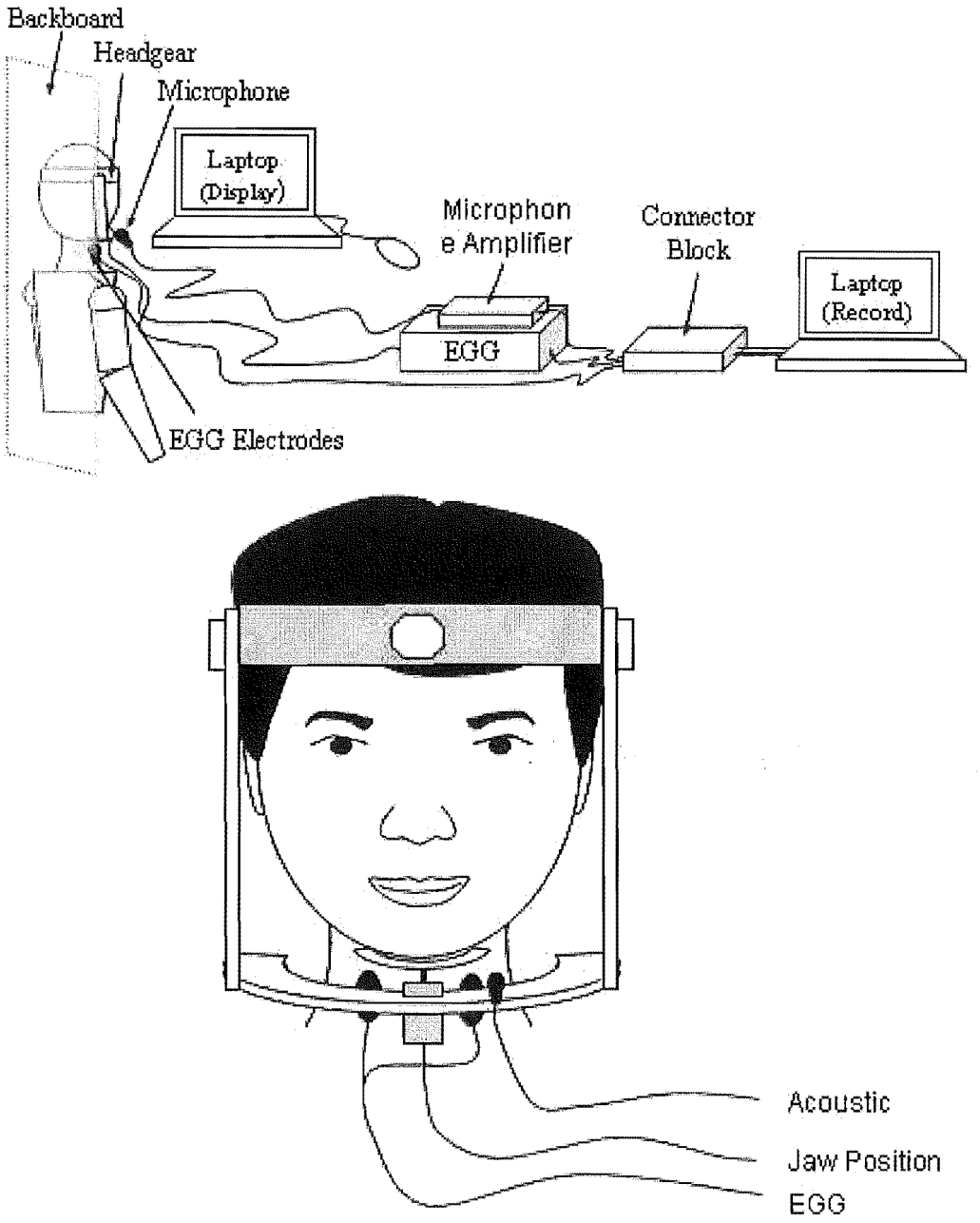

Figure 6.2 The diagram of the instrumentation setup for the experiment (top) and close-up front-view of a subject ready to begin the experiment (bottom).

6.2.3 Procedure

During the experiment, the head mount with the jaw position sensor was placed on the subject's head. The microphone attached to the head mount was placed off-axis at a distance of approximately 5 cm from the subject's mouth. The two electrodes on the electroglottograph were placed on either side of the subject's thyroid cartilage and held in place by an elastic band. The subject was asked to sit in the upright position with the Velcro strapped back-end of the head mount held against a rigid backboard to minimise head movement during recording. The subject was also asked to close their jaw before and after each sentence was read so that the neutral point of the jaw position could be recorded.

A second laptop computer (Acer 450MHz Pentium 3) was placed in front of the subject with a display (Power Point 2000) showing the sentence to be read. The experimenter controlled the slide rate of the Power Point display and asked the subject to say each sentence. The order of the speech task was randomised to avoid any presumption of what the next word was going to be before the target was displayed on the screen. The simultaneous acoustic, jaw position, and EGG recording system was activated by the experimenter pressing the "Enter" key on the recording computer before the start of each set of five sentences. Each set of five sentences had in between them approximately one-second pause, which was achieved by the experimenter manually delaying the display of the next sentence on the screen.

6.2.4 Data Analysis

In total, 650 utterances (13 contexts x 5 vowels x 10 subjects) were recorded. Figure 6.3 shows a sample of a sequence of 5 utterances. For each utterance, three glottal parameters (*F0*, *OQ*, *SQ*) were measured from the EGG signals and the magnitude of the jaw opening was derived from the jaw position signal. For signal segmentation, all three simultaneously recorded signals were displayed on a computer monitor using the MATLAB12 Signal Processing Toolbox. The experimenter listened to the acoustic signal and at the same time visually selected from each utterance the target vowel segment in the electroglottogram showing a clear periodic pattern. Among all the 650 signals segmented, the number of cycles chosen in a segment ranged from 2 to 38 (Mean = 11 cycles, SD = 6).

Fundamental Frequency

The average *F0* was obtained from the EGG signal by dividing the number of cycles in the selected vowel segment by the duration of the whole segment (in seconds), making sure that the segment started and terminated at a trough of the EGG signal.

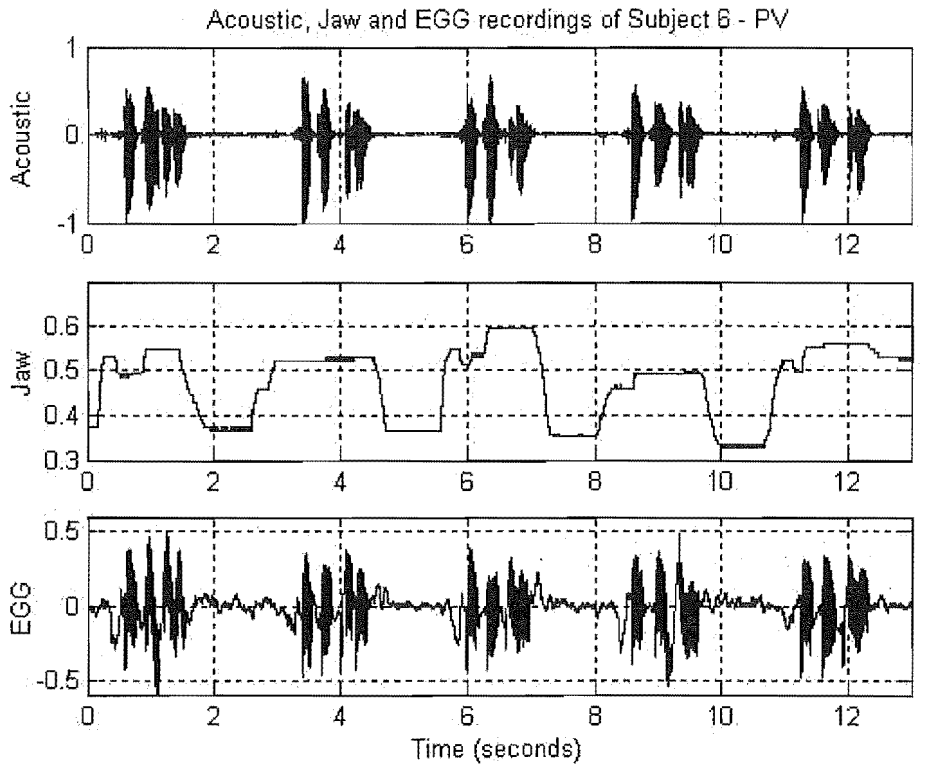

Figure 6.3 Acoustic signal (top), jaw position signal (middle) and EGG signal (bottom) for the 5 phrases: "Say po again", "Say pe again", "Say pa again", "Say pu again" and "Say pi again".

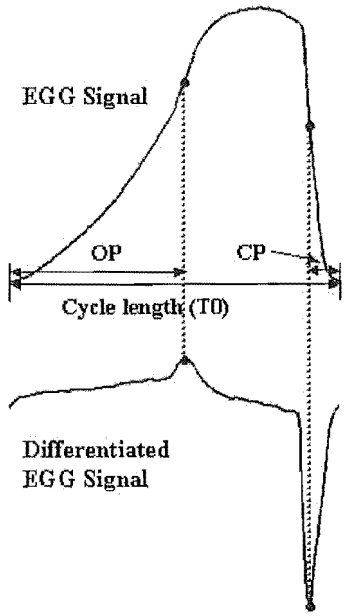

Figure 6.4 An example of a stylised EGG waveform and its time derivative waveform showing how *T0*, *OP* and *CP* were obtained.

Speed Quotient and Open Quotient

In the stylised electroglottogram shown in the top half of Figure 6.4, a positive going EGG signal represents a decrease in vocal fold contact. Once the segment of interest had been selected, the average waveform was obtained by aligning each cycle of the waveform along its falling edge and taking the mean of each sample instant over the set of cycles in the segment. As shown in Figure 6.4, the beginning and the end of the average waveform were marked as the beginning of the opening phase (*OP*) and the end of the closing phase (*CP*), respectively. The average waveform was then differentiated. As also shown in Figure 6.4, the positive peak and negative peak of the differentiated waveform were marked as the end of the *OP* and the onset of the *CP*, respectively. Speed quotient and open quotient are defined in Eqn.6.1 and Eqn.6.2 as follows:

$$SpeedQuotient, SQ = \frac{OP}{CP} \tag{6.1}$$

$$OpenedQuotient, OQ = \frac{OP}{T0} \tag{6.2}$$

where *T0* is the cycle length (*1/F0*). The algorithm used for determining *OP* and *CP* is similar to the LF model [FLL85] used in the study of glottal airflow.

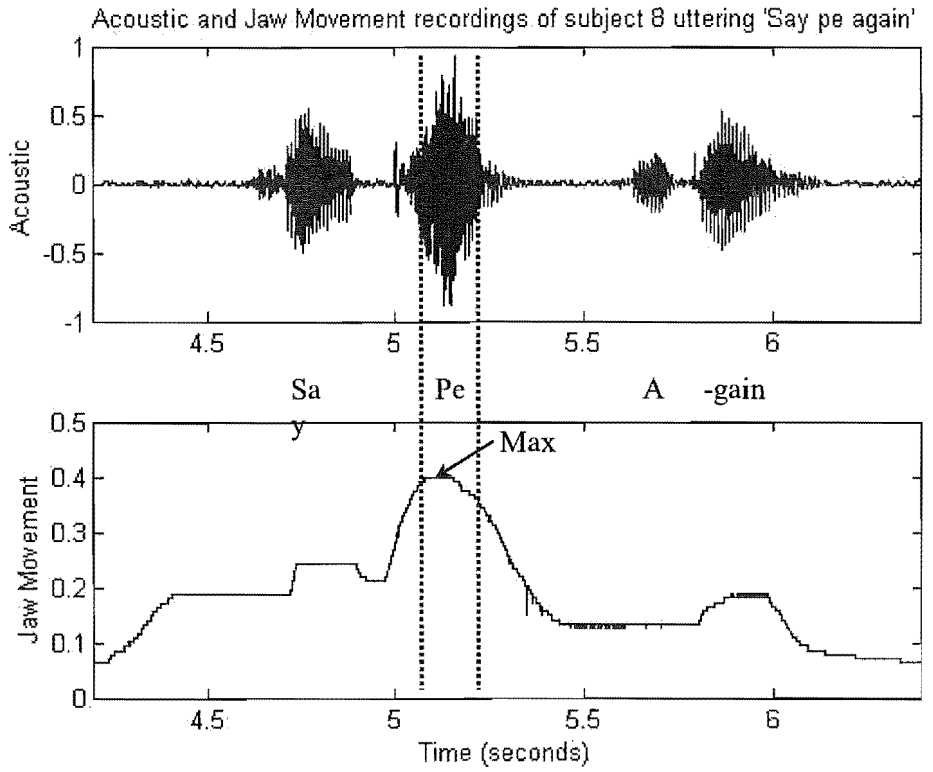

Figure 6.5 An example of a single utterance with acoustic and jaw movement signals, showing how the magnitude of jaw opening was extracted.

Magnitude of Jaw Opening

With higher output voltage from the sensor corresponding to larger jaw opening, the trace obtained via the recording system of the jaw position was used to extract the magnitude of jaw opening during speech production. Within the same time frame as selected for acoustic and EGG signal analysis, the measure of the magnitude of jaw opening was taken at the time of maximum jaw opening during the vowel portion within that utterance (Figure 6.5).

Data Normalisation

The range of *F0* and the magnitude of jaw opening varied across speakers. Therefore, the *F0* and the magnitude of jaw opening data were normalised for each subject to control for inter-subject variation for better determination of the vowel effect on these measures. The normalisation of *F0* was performed by first calculating the mean and standard deviation of an individual's *F0* over all 5 vowels in the same context. The normalised score (Z-score) for each *F0* measure was obtained by subtracting the mean from each measure and then dividing the remainder by the standard deviation (Table 6.1). The same operation was performed for the magnitude of jaw opening.

	Raw Data for F0 (Hz)					Raw Data Mean	Normalised Data (F0)				
	a	e	I	O	u	Mean (SD)	a	e	I	o	u
1	133	132	144	127	143	135.8 (7.40)	-0.38	-0.51	1.11	-1.19	0.97
2	116	126	129	120	129	124 (5.79)	-1.38	0.35	0.86	-0.69	0.86
3	116	124	133	121	135	125.8 (8.04)	-1.22	-0.22	0.90	-0.60	1.14
4	121	129	132	124	139	129 (7.04)	-1.14	0.00	0.43	-0.71	1.42
5	127	137	141	131	138	134.8 (5.67)	-1.37	0.39	1.09	-0.67	0.56
6	131	134	141	142	156	140.8 (9.68)	-1.01	-0.70	0.02	0.12	1.57
7	138	133	151	138	146	141.2 (7.19)	-0.45	-1.14	1.36	-0.45	0.67
8	133	129	138	122	134	131.2 (6.06)	0.30	-0.36	1.12	-1.52	0.46
9	125	121	144	122	136	129.6 (10.01)	-0.46	-0.86	1.44	-0.76	0.64
10	135	136	136	129	137	134.6 (3.21)	0.12	0.44	0.44	-1.74	0.75
11	126	126	132	126	137	129.4 (4.98)	-0.68	-0.68	0.52	-0.68	1.53
12	128	124	130	121	141	128.8 (7.66)	-0.10	-0.63	0.16	-1.02	1.59
13	125	132	132	131	141	132.2 (5.72)	-1.26	-0.03	-0.03	-0.21	1.54
						Overall Mean:	**-0.72**	**-0.24**	**0.67**	**-0.81**	**1.09**

Table 6.1 Example normalisation values of *F0* for one subject.

6.2.5 Statistical Analysis

For each subject, all measures were averaged over all 13 trials for each vowel. For the measure of the magnitude of jaw opening, a total of 35 data points (7 subjects 5 vowels) were obtained. The recordings for three subjects (subjects 5, 7, and 10) were excluded because there was clear evidence that jaw opening had exceeded the measurable range. For each of the measures of fundamental frequency, open quotient, and speed quotient, a total of 40 measurements (8 subjects 5 vowels, averaging over all 13 contexts) were obtained. The EGG signals were of insufficient quality to allow estimations of parameters in two subjects (subjects 2 and 4).

All data were submitted to a series of one-way Repeated Measures (RM) analysis of variances (ANOVAs) to determine whether there is a vowel effect on these measures. A series of Pearson product moment correlations were also performed to determine whether *F0*, *SQ*, *OQ*, and magnitude of jaw opening are correlated with one another.

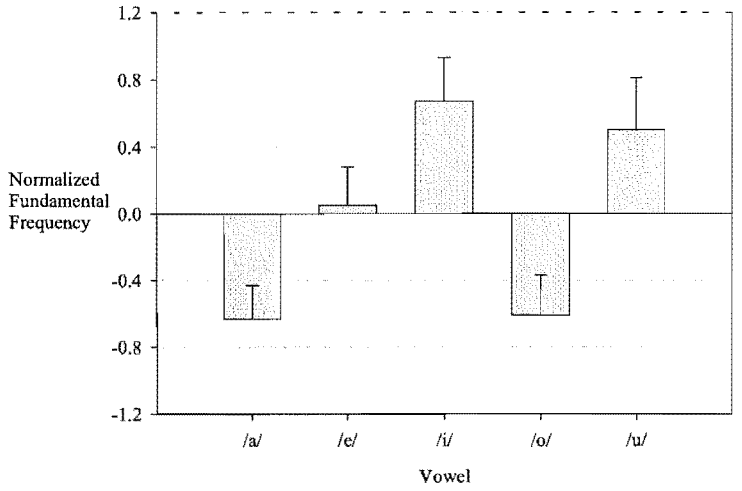

Figure 6.6 Means and standard deviations of normalised *F0* across vowels (negative value implies lower *F0*).

6.2.6 Results

Results of a series of one-way RM ANOVAs revealed a significant vowel effect on the magnitude of jaw opening [$F(4,24) = 25.512$, $p < 0.001$], fundamental frequency [$F(4,28) = 45.415$, $p < 0.001$], and speed quotient [$F(4, 28) = 5.233$, $p < 0.003$], but no significant vowel effect on the measure of open quotient [$F(4, 28) = 0.501$, $p < 0.735$].

Fundamental Frequency

Vowel *F0* range for this experiment before data normalisation was found to be between 94-254Hz with an average of 166Hz. The intra-subject difference between the highest vowel *F0* and lowest vowel *F0* is between 4 to 45Hz (with a SD ranging from 0.28 semitones to 1.9 semitones). Figure 6.6 shows that /i/ has the highest *F0*, followed in order by /u/, /e/, /o/, and /a/. Results of post-hoc pairwise multiple comparison procedures with the Student-Newman-Keuls Method indicated that all comparisons between vowels on the measure of *F0* were significant ($p < 0.05$) except for the comparisons between /i/ and /u/ and between /o/ and /a/.

Magnitude of Jaw Opening

Results of post-hoc pairwise multiple comparison procedures revealed that all comparisons between vowels on the measure of the magnitude of jaw opening were significant ($p < 0.05$) except for those

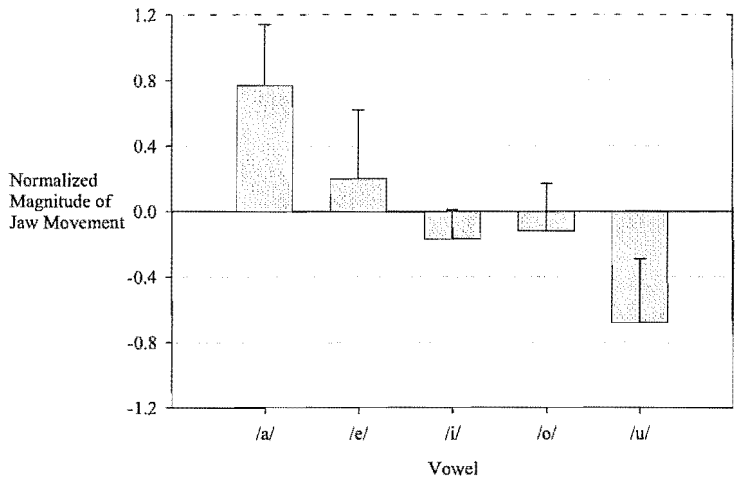

Figure 6.7 Means and standard deviations of normalised magnitude of jaw opening across vowels (negative value implies smaller jaw opening).

among /e/, /o/, and /i/. As shown in Figure 6.7, the vowel /a/ in average exhibited the greatest magnitude of jaw openings, followed in order by /e/, /o/, /i/, and /u/.

Speed Quotient and Open Quotient

Results of post-hoc pairwise multiple comparison procedures indicated that only /u/ differed significantly from the rest of the vowels on the measure of speed quotient ($p < 0.05$). Figure 6.8 shows that /u/ has the lowest speed quotient. Figure 6.9 shows that the average measures of open quotient remain relatively constant across vowels.

Relationships between *F0* and Other Measures

Results of a series of Pearson Product Moment Correlations conducted to determine the relationships between *F0* and the other three experimental measures (i.e., *SQ*, *OQ*, magnitude of jaw opening) revealed a significant negative correlation between *F0* and the magnitude of jaw opening ($r = -0.624$, $n = 35$, $p = 0.0009$) but no significant correlation between *F0* and *OQ* ($r = -0.109$, $n = 25$, $p = 0.605$) or between *F0* and *SQ* ($r = -0.235$, $n = 35$, $p = 0.259$). The inverse relationship between *F0* and the magnitude of jaw opening can be observed in Figure 6.10.

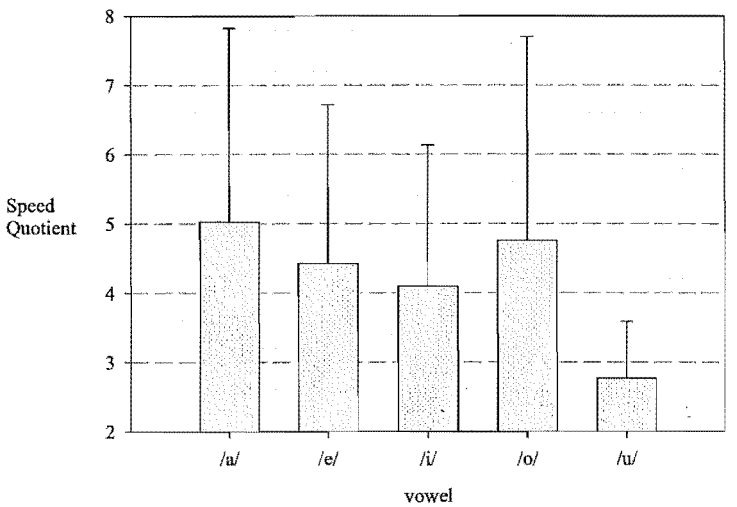

Figure 6.8 Means and standard deviations of *SQ* across vowels.

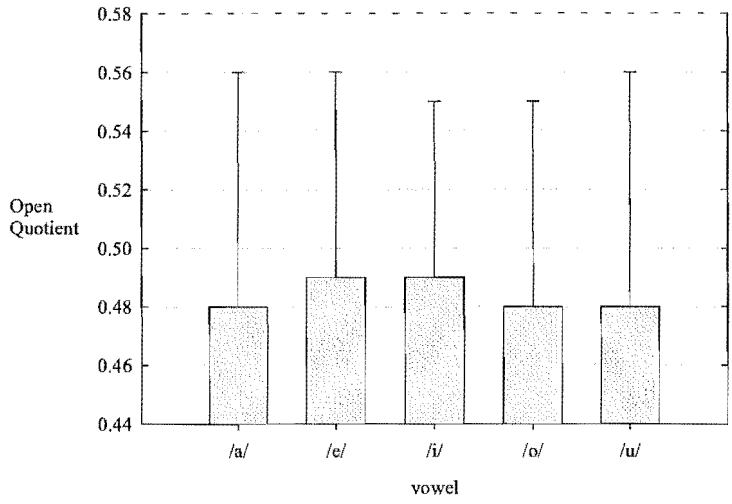

Figure 6.9 Means and standard deviations of *OQ* across vowels.

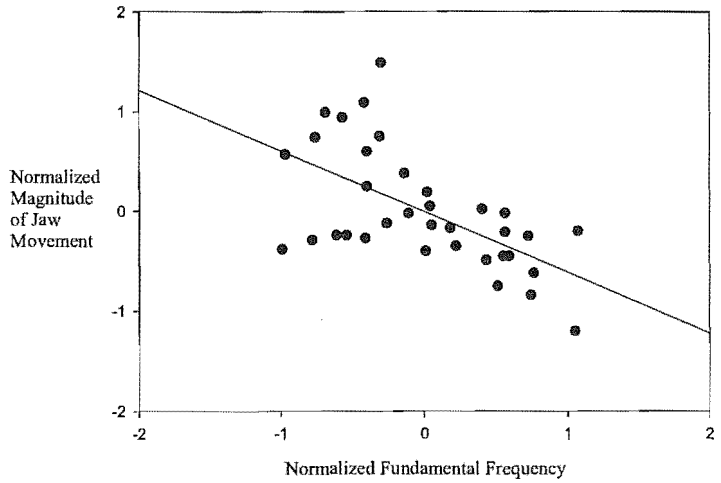

Figure 6.10 Frequency versus jaw position for 7 subjects (excluding subjects 5, 7 and 10).

6.2.7 Discussion

The vowel measurement did not appear to vary by context (V, CV and VC), possibly due to the fact that only the stable segment of the vowel was extracted from the EGG and jaw signals for analysis. Therefore, data derived from V, CV and VC syllables were combined together for further statistical analysis.

The aim of this study was to determine whether the magnitude of jaw opening was related to the vibratory pattern of vocal folds in neutral (non-emphasised) speech. Based on our observations of simultaneously recorded jaw magnitude and glottal parameters in different vowel production, we did find the magnitude of jaw opening to be inversely related to the *F0* of the vocal fold vibration. Vowel identity was found to affect the magnitude of jaw opening, *F0*, and *SQ*. These findings confirmed our hypothesis that the magnitudes of jaw opening in different vowel productions are associated with changes in voicing pattern. This suggests that the jaw signal may be used to control the *F0* of an artificial larynx to improve the voice source simulation.

Our finding that *F0* and the magnitude of jaw opening were inversely related is consistent with Marasek's [Mar96] conclusion that opened vowels had lower *F0* than closed vowels and with Zawadzki and Gilbert's [ZG89] finding that *F0* is related to the position of the mandible. We found that a significant *F0* difference existed among three clusters of vowels, namely /i/, /u/, /e/, and /o/, /a/ while the magnitude of jaw opening significantly differed among three clusters of vowels, namely /u/, /i/, /e/, /o/, and /a/. These findings suggest that three *F0* selections can be designated for three ranges of magnitude of jaw opening, namely, high *F0* for small jaw movement /u/, /i/, moderate *F0* for moderate jaw movement /e/, and low *F0* for large jaw movement /o/, /a/. The implication for designing an artificial larynx is that the device can be customised to an individual's vowel-related jaw movement.

The intra-subject difference between highest vowel *F0* and lowest vowel *F0* of 4-45Hz (or SD of 0.28-1.9 semitones) in our study is slightly wider than the values reported in the literature [Pet78, WL95]. Although the intra-subject vowel *F0* range is not as high as that in running speech (3.4 semitones), it still constitutes a significant portion of the change in pitch. For a jaw tracking voice synthesiser, *F0* scaling across the vowels might be required to bring the intrinsic *F0* range up to the normal conversation speech range.

The measures of *OQ* and *SQ* did not appear to be affected by vowel identity or the magnitude of jaw opening except for the finding that /u/ has a significantly lower *SQ* than all other vowels. Our finding that *OQ* did not vary by vowel or the magnitude of jaw opening agreed with Marasek's study [Mar96]. However, our findings regarding *SQ* are different from those reported in the study. In Marasek's study [Mar96], which compared /a/, /e/, /i/, /o/, and /u/, the female data (5 females) showed that only /o/ had a significantly lower *SQ* than all the other vowels while the male data (5

males) showed a constant *SQ* at 1.26. Since both /o/ and /u/ involve lip-rounding, it is speculated that lip-rounding may be associated with a more abrupt vocal fold opening and thus a lower *SQ*.

Since jaw movement is not limited to a single plane [DH02], a number of investigators have used multiple sensors for jaw tracking in their studies. In an experiment conducted by Erickson *et al* [EIEF04], both vertical and horizontal jaw movement were measured. Of the two subjects studied, one showed that the change in vertical position of jaw is more pronounced than that of the horizontal jaw position while the other showed that the reverse is true, although in the latter case, the same vowel /a/ was used but at a different tone level (emphasis). Another investigator [Geu01] used 3 sensors to track the jaw movement of their subjects but only presented jaw height in their results. There are also studies where single sensor jaw tracking is used (e.g. [EFP98]). In this study, we find that vertical jaw movement tracking with one sensor is sufficient for the purpose of our experiment. The range of the sensor, however, is not wide enough as apparent in the experiment where data from three subjects had to be excluded because the jaw opening for these subjects exceed the detector's measurable range. A non-contact sensor with a working range of 25mm is required. A sensor based on light reflection may be appropriate to minimise problems with friction impeding the natural jaw movement.

Without an anti-aliasing filter prior to digitisation, signals at high frequencies can be aliased. However, since the EGG spectrum has about 12dB/oct attenuation, at 2 kHz (Nyquist frequency) the signal amplitude has dropped by approximately 50dB from the initial *F0* amplitude. Thus, frequencies that fall beyond the Nyquist frequency become less relevant in the case of the EGG signals. In the study the /e/ and /i/ acoustic signals were sometimes difficult to differentiate perceptually. This is because the second formant for these two vowels falls beyond the Nyquist frequency. Because a record of the order of utterance by each subject was available, the potential confusion was avoided. It would however be better to increase the sampling rate in future experiments to prevent such confusion.

Our study is limited in its generalisation as only male subjects were included. Further studies on female adults as well as other age groups are needed to determine whether the relationship between the magnitude of jaw opening and glottal parameters found in this study applies to other populations. Inclusion of different modes of phonation, more contextual variability, and a greater variety of speech tasks in the sampling of phonatory behaviours would also be useful for clarifying the extent of the physiological linkage between jaw opening and the voice source.

6.2.8 Conclusions

This study has demonstrated that *F0* increases as the magnitude of jaw opening decreases in vowel production. Changes of *OQ* and *SQ* obtained from the EGG signal did not appear to be affected by vowel identity or the magnitude of jaw opening except that the vowel /u/ is associated with a

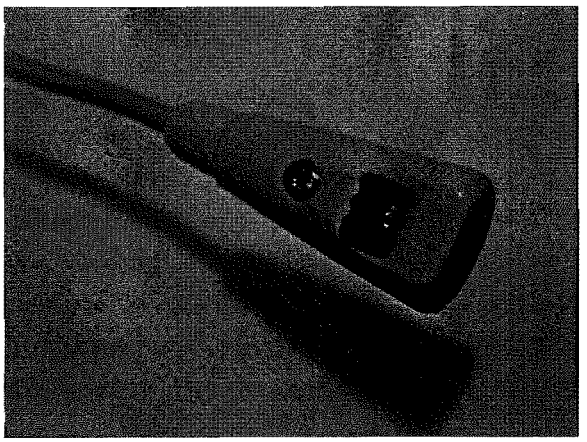

Figure 6.11 The reflective object sensor used for measuring jaw movement.

lower *SQ*. The findings indicate that long-term *F0* variation in artificial voice production may be implemented through a jaw tracking scheme for better source simulation. Jaw tracking is therefore incorporated in the design of the prototype artificial speech device (*MyVoice2*) in Chapter 8.

6.3 Jaw movement detector for prototype development

The two main problems encountered with the mechanical jaw movement detector in the experiment above is that the jaw opening of some subjects exceed the detector's measurable range and that friction from the potentiometer may affect the natural jaw movement. In order to overcome these problems, a non-contact detector using reflective object sensor like the one shown in Figure 6.11 has been introduced. This sensor uses light reflection technology to measure the distance of an object (in this case, the tip of the chin) from the surface of the sensor. The reflective object sensor has a working range of >25mm, which meets the requirements for measuring the jaw opening of most people during normal speech. This sensor is used for the prototype artificial speech device in Chapter 8.

Chapter 7

Contribution of glottal wave shape to natural sounding voiced speech

The glottal pulse is the natural voice source generated by the vibrating vocal folds during voicing. Studies from other investigators suggest that glottal pulse shape affects the naturalness of speech [Ros71, CW90, Hol73], as do pitch variation and jitter [CC96]. Over the years, a number of glottal source models have been used to improve the quality of synthesised speech including the LF-model [FLL85], single-mass model [FL68, Luc04], two-mass model[IF72], multi-mass model [KAR99, Tit73, Tit74, TS75], finite-element model[ABT00], impulse [CW90], and inverse filtered glottal waveforms [Ros71, Hol73], with varying results.

The inverse filtered glottal waveform obtained from the audio recordings of an original speaker was found to give one of the best results as a glottal source as it retains many of the qualities of the original speaker [HHKM99]. Results from inverse filtered acoustic waveforms suggests that the glottal sound source for the modal register has a shape similar to that of a skewed sinusoid, with an 8-12 dB/oct attenuation (slope) in the frequency spectrum [Pic98]. However, the inverse filtered glottal waveform is difficult to estimate and requires a high quality microphone with good low frequency response [CW90].

Electroglottography (EGG) is a device that provides information about the vocal folds contact by measuring the electrical impedance between 2 electrodes placed on either side of the thyroid cartilage [Chi84]. The EGG signal obtained from normal subjects displays an 8-12dB/oct attenuation in its frequency domain representation, similar to that of the glottal airflow waveform mentioned above. Since the EGG signal can be measured directly, in contrast to the glottal sound source which can only practically be derived from inverse filtering of the acoustic signal, it is convenient to use a model derived from the EGG signal as an alternative glottal sound source for an artificial larynx. The EGG signal is also independent of the vocal tract shape (e.g. vowel independent), making the

design of the glottal model simpler.

The first section of the chapter presents a study of the glottal pulse shape by measuring the EGG signal in the speaking and singing range. This is to allow for the change in voice modes (if they occur) so that different mode of voicing can also be included in the study. It is envisaged that the information gathered from this experiment may be used to change the glottal pulse shape dynamically as *F0* varies to improve the naturalness of artificial speech. The second section is on voice synthesis where 3 glottal models and a vocal tract model discussed in Chapter 5 were used for the artificial voice synthesis. The synthesised voice in this section is used for the final section where the identity and quality of artificial voice generated can be compared.

7.1 Waveform shape experiment

The purpose of this experiment is to establish that glottal waveform shape changes with pitch and to show that the information gathered from the experiment can be used to design a mathematical glottal pulse model, the twin-bar model, discussed in Chapter 5.

7.1.1 Setup

The voice recordings were carried out in an acoustically tiled quiet room. As the EGG signal does not contain enough information to reliably recognise vowels, the acoustic signal was also recorded to verify that the vowel sounds uttered by the subjects were consistent with the directions given. The acoustic signal was also used to determine the subject's voice register.

Acoustic and EGG signals were recorded simultaneously. For the acoustic recording, a mono headset with boom microphone (Labtec Axis-501, manufactured by Logitech Inc. USA) was used to generate a tone in the subject's ear and to record the subject's voice respectively. The microphone (positioned to the left of the mouth approximately 5cm from the lips) was connected to an external microphone amplifier (Eurorack MX602A from Behringer, Germany), the output of which was digitised. A Kay Elemetrics Model 6103 was used to measure the EGG signal. This device measures the electrical resistance between two electrodes placed on either side of the thyroid cartilage (see Figure 6.2 for the location of the EGG electrodes on the neck). The signal obtained is the unfiltered EGG waveform. The unfiltered EGG waveform was inverted so that opening of the vocal folds corresponds to a positive change in signal amplitude (see Figure 7.1).

Before digitising, the signals were anti-alias filtered at 5kHz. Digitising was performed with a 650MHz Pentium 4 laptop (Compaq, Taiwan) via a 12-bit A/D converter (DAQCard-AI-16E-4 from National Instrument, USA) at a sampling rate of 22 kHz and with 16-bit resolution. The digitised signals were then stored in the laptop in .wav format. The data acquisition software and waveform analysis software written in MATLAB12 were used for signal acquisition and data retrieval.

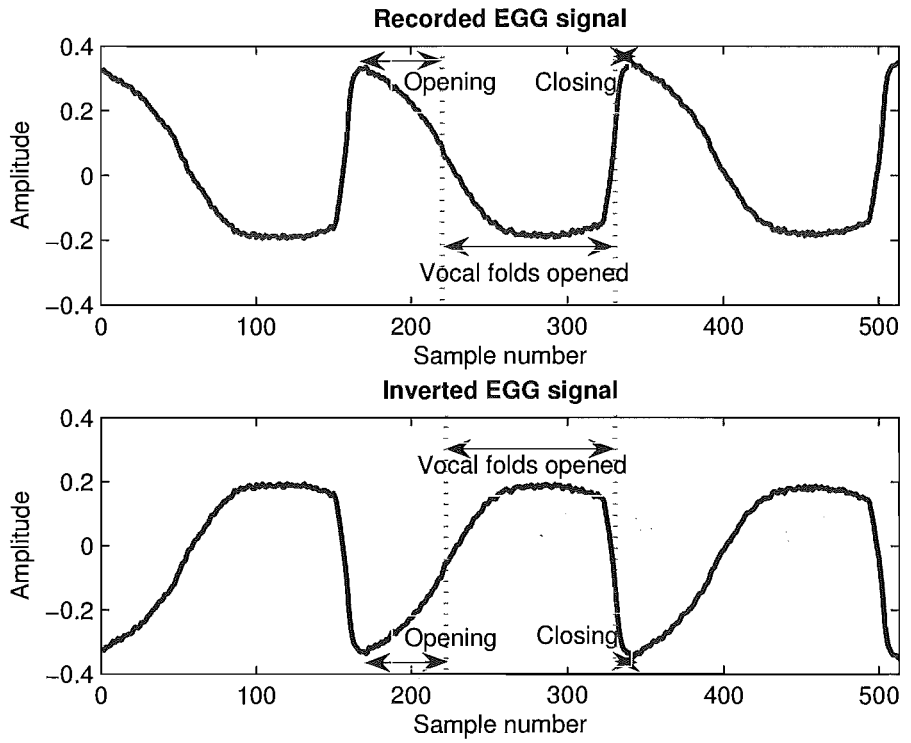

Figure 7.1 The recorded EGG signal (top) and the inverted EGG signal that is used for EGG analysis (bottom).

The Visi-Pitch system (Model 6087AT, KAY Elemetric Corp., New Jersey, USA) was used to extract the habitual pitch of each subject at the beginning of the experiment. For the rest of the experiment the Computerized Speech Lab (CSL, New Jersey, USA) Voice Range Profile program was used to provide visual feedback to guide the subjects in maintaining the required pitch and loudness. A Pentium 3 laptop (Acer) was used to display the visual feedback.

7.1.2 Procedure

The subjects for this study consisted of 14 adult New Zealanders of European descent with English as their first language and with no history of speech and hearing disorders or apparent voice and resonance problems. All subjects were non-smokers. One subject was a trained opera singer, 4 subjects had some form of singing experience and 9 subjects had no singing experience at all. Subjects included 6 males and 8 females aged from 20-52 (Mean (M) = 30.67 years, SD (M) = 11.86 and Mean (F) = 31.25, SD (F) = 9.00). One additional female subject was excluded from the study because adequate quality EGG signals were not able to be recorded.

During the experiment, the subjects were asked to perform a number of tasks. The first task was designed to enable the individual's habitual pitch to be determined. The second and third tasks were designed to enable the EGG pulse shape to be estimated in the speaking range and singing range respectively.

Task1: The subject was asked to read a standard passage (the first sentence of the rainbow passage [Fai60]). The acoustic signal recorded during task 1 was then used to extract the habitual pitch of the subject before proceeding to tasks two and three.

Task2: The speaking range test: the subjects were asked to listen to a tone and to try to match the given tone using different vowel sounds (e.g. /a/, /e/, /i/, /o/, /u/ and /er/). Nine tones were given for each vowel sound at the subject's habitual pitch and at each of the 4 semitones above and below. The number of semitones from the habitual pitch is given by

$$n = 39.86 \log_{10} \frac{F1}{F0}, \tag{7.1}$$

where $F0$ is the habitual pitch and $F1$ is the pitch of interest. At the same time, the subjects were asked to use the display on a computer screen as visual feedback to maintain their loudness and pitch. At the beginning of each session, a 2-minute-long test run was carried out in order to familiarise the subject with the visual feedback program. The subject's ability to match a given tone exactly was not particularly important as it was given only to act as a guide for the subject to produce different

pitch levels that were within ±4 semitones of their speaking range.

Task3: For the final task, starting from the subject's habitual pitch and for each of the vowel sounds stated above, the subject was asked to sing an octave major scale in descending order and then in ascending order. The tone for habitual pitch was played on the headphones so that the subject knew where to begin. For Task2 and Task3, each sound was required to be sustained for at least 1 second.

A second trial of 4 sounds randomly chosen from Task2 was also recorded and compared with their corresponding sounds in the first trial to test the reliability of the recorded data.

7.1.3 Waveform analysis

A total of 2156 utterances (14 x (6x9 + 6x16 + 4)) were recorded in this experiment. A pre-analysis selection was carried out to remove waveforms that were too weak (small SNR) or too unstable (signal amplitude that varied extensively through an utterance) to be analysed. The pre-analysis selection was done by visual inspection of each recorded waveform. Eventually only 1901 utterances were found to be useful for analysis. The region of interest for each of the EGG waveforms was then selected manually (usually in the middle of an utterance where the waveform was stable). The duration of the waveform of interest varied from 0.1s - 2s depending on the recorded data.

An algorithm written in MATLAB was used to extract the parameters from each selected segment. First, the algorithm found the minimum points of each cycle of the EGG waveform and then removed the baseline shift of the waveform by subtracting from each point on the EGG signal a value equal to the linear interpolation between the nearest two minimum points. This correction is illustrated in Figure 7.2.

The average period of the EGG cycle was calculated and any EGG cycles with periods outside ±2.5% of the average period were discarded. This step was important as the computed mean pulse shape may have been distorted if EGG waveforms with a large variation in period were included.

The algorithm then separated the baseline compensated waveform into individual cycles and aligned each cycle at the instant of closure (maximum negative slope, *sCP*, Figure 7.3). The mean waveform was computed from the aligned individual cycles. Glottal parameters were obtained from the mean waveform.

For each utterance, 4 glottal parameters (*T0*, *OP*, *CP* and *sCP*) were measured from the mean EGG signal and its derivative (*DEGG*). Refer to Figure 7.3 for the quantities mentioned. The values *OQ*,

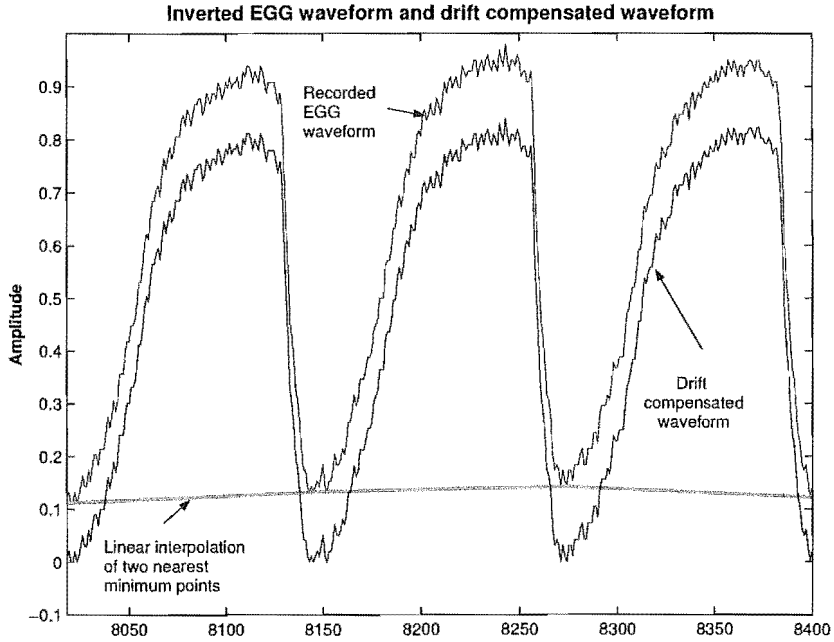

Figure 7.2 Inverted EGG waveform and the same waveform with baseline shift removed (same as Figure 4.8).

Table 7.1 Comparison of two sets of 56 sounds from 14 subjects.

	F0 (Hz)	*OP* (sample number)	*CP* (sample number)	*sCP* (amplitude/sample)
Maximum Difference	68	42	13	0.024
Mean Difference	7	9	2	0.004
Standard Deviation	11	9	3	0.005

SQ (see Eq.4.1 and Eq.4.2) and *F0* were then derived from these parameters. The pitch period (*T0*), measured in seconds is the duration between two successive negative peaks on the *DEGG* waveform. The fundamental frequency (*F0*) is the inverse of *T0*.

7.1.4 Results and discussion

Reliability of the Waveform Analysis Method

To test the reliability of the parameter extraction program, for each subject 4 sounds randomly chosen from Task 2 were repeated. The parameters (*F0*, *OP*, *CP* and *sCP*) of the two sets of 56 sounds (4 sounds × 14 subjects = 56 samples) were compared.

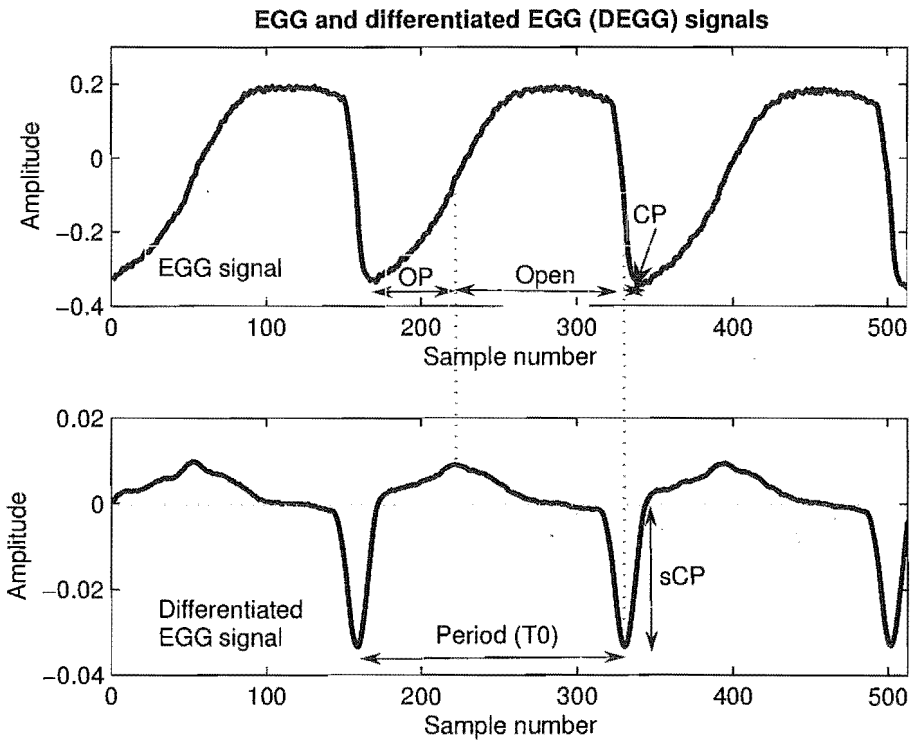

Figure 7.3 An example of an inverted EGG waveform and its time derivative waveform showing how *T0*, *OP*, *CP*, *Open* and *sCP* are obtained.

Table 7.2 Comparison of two sets of 4 sounds from subject 16, a singer.

	F0 (Hz)	*OP* (sample number)	*CP* (sample number)	*sCP* (amplitude/sample)
Maximum Difference	1	11	1	0.004
Mean Difference	1	9	0	0.002
Standard Deviation	1	4	1	0.002

Table 7.1 shows the maximum difference, mean difference and standard deviation of the two sets of data from all subjects. The main reason for the large variation for all parameters is because most subjects were not trained signers and therefore were unable to reliably repeat a sound that matched the given pitch.

In order to reduce the variability, data from subject 16, a trained opera singer, was analysed separately (see Table 7.2). Results from the data by this subject showed marked reduction in the differences between the two sets of parameters. This indicates that the analysis method is repeatable.

Habitual Pitch

The average habitual pitch for the male subjects was 116.7 Hz (97.4 - 146.6 Hz) while the average habitual pitch for the female counterpart was 193.9 Hz (155.7 - 222.5 Hz). These figures are consistent with the results reported in the literature [CC96, BYNG98].

Vowel Effect

Figure 7.4 shows the mean EGG waveforms for subject 9 uttering 6 different vowels across the subject's speaking range which is between 88.2 and 135.3 Hz (or between -3.3 and 4.0 semitones from his habitual pitch at 107.1 Hz). It can be observed that vowel effect on the EGG pulse shape is small. The slight variation in pulse shape from vowel-to-vowel could possibly be due to the orientation of the articulatory structures (e.g. tongue, jaw and lips) for different vowels which may cause the larynx to shift. As the change in pulse shape is quite small (mainly concentrated near the tip of the waveform), it should not affect the parameter extraction in the waveform analysis. The skewness change at lower frequencies in each plot probably indicates that the vocal folds were starting to strain or starting to shift to a different vibrating mode. Similar observations were made on the other subjects.

In Figure 7.5 the opened quotient (OQ) and speed quotient (SQ) are plotted versus $F0$ for Subject 9 for the 6 vowels. The vowel effect on OQ seems to be minimal; all plots exhibit a slight increase in OQ with $F0$. For SQ all vowels exhibit a decrease of SQ with $F0$ except for /u/; in this case no significant change is observed. Again, the results shown in Figure 7.5 are representative of the

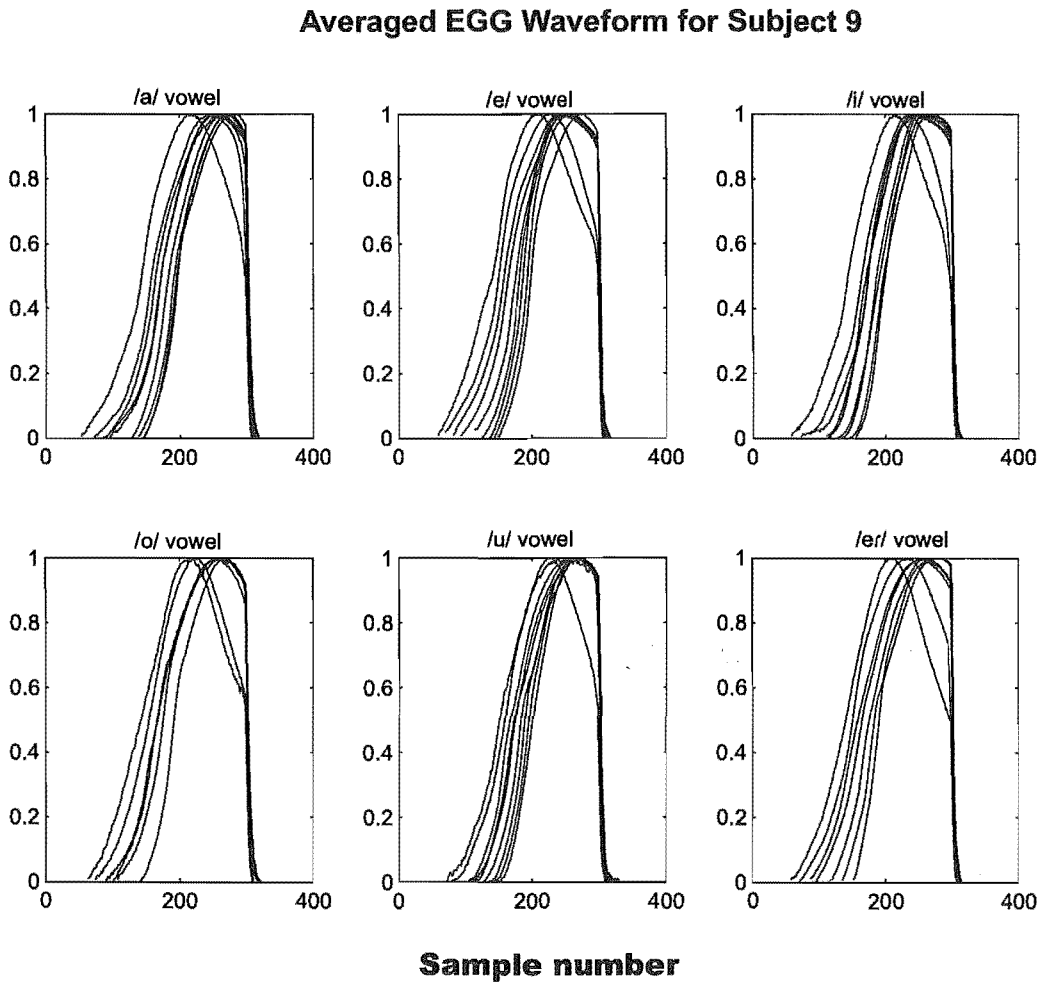

Figure 7.4 The mean EGG waveforms of Subject 9 for separate vowels across the speaking range.

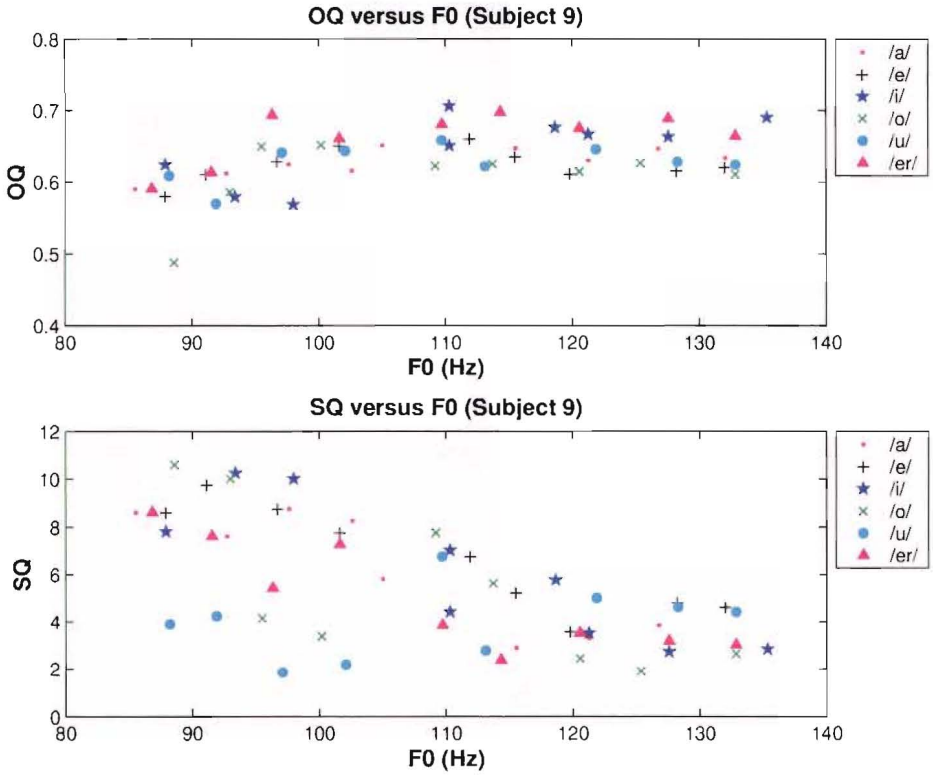

Figure 7.5 Open quotient and speed quotient versus fundamental frequency plots for subject 9.

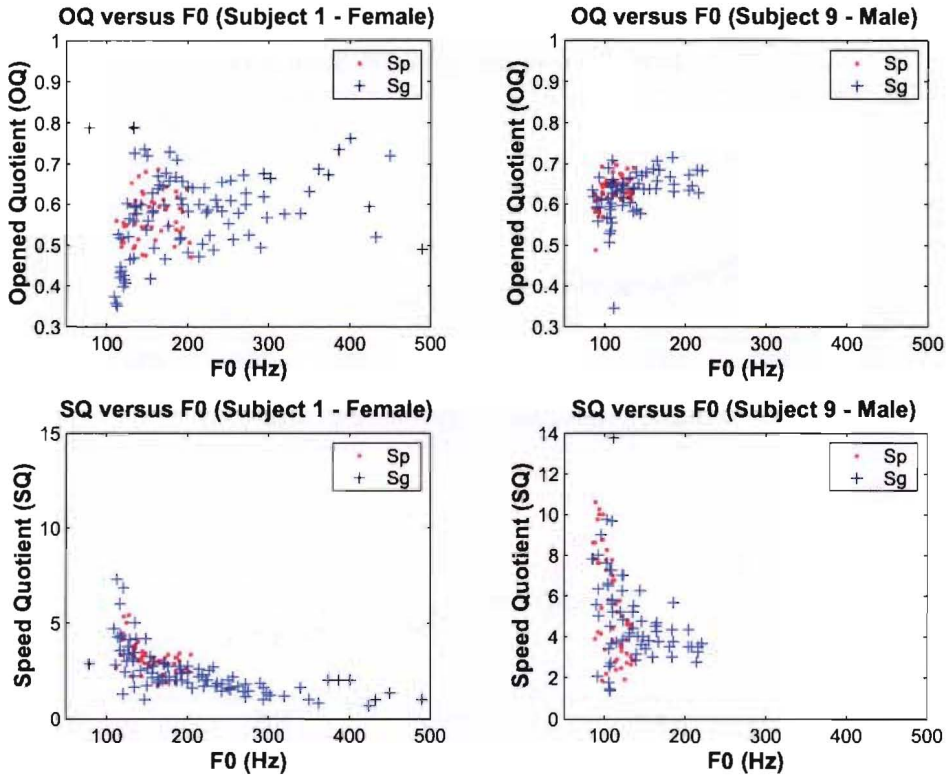

Figure 7.6 Open quotient and speed quotient versus fundamental frequency plots for Subject 1 and Subject 9.

observations in all subjects.

Speaking Range and Singing Range

The plots of OQ and SQ values versus frequency for Subject 1 and Subject 9 in Figure 7.6 show that there seems to be no significant difference in these parameters between speaking range values and singing range. This implies that the data from the speaking range and singing range can be combined together for analysis purposes.

The relationships between OQ and $F0$, and between SQ and $F0$, can be found by linear regression over all subjects. They are:

$$\ln OQ = 0.136 \ln F0 - 1.19 \tag{7.2}$$

$$\ln SQ = -0.85 \ln F0 + 5.48 \tag{7.3}$$

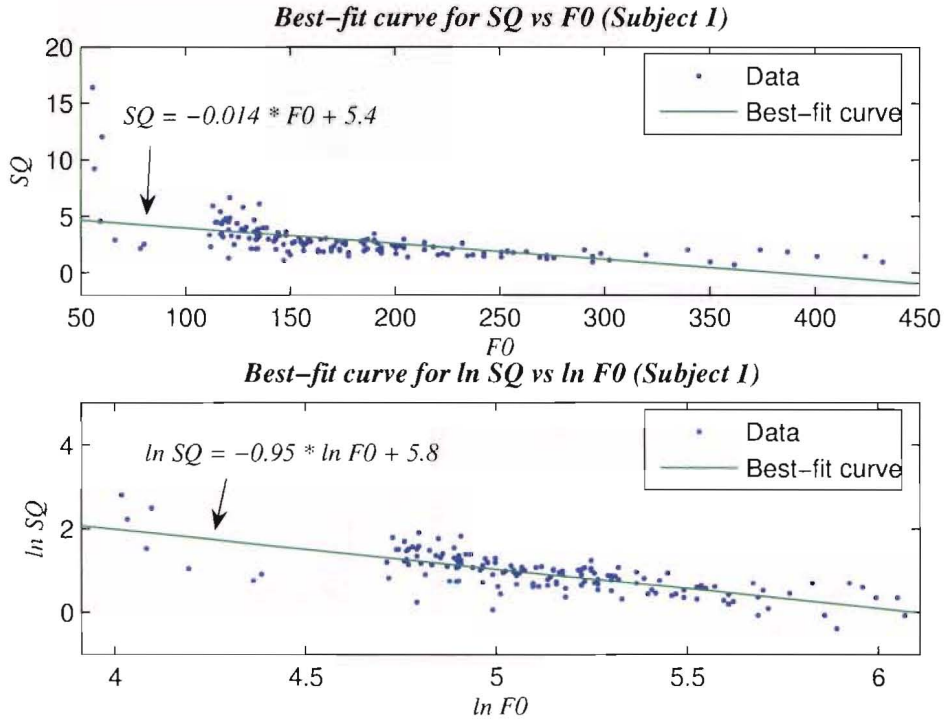

Figure 7.7 The best-fit lines for (a). *F0* vs *CP* (linear scale) and (b). ln *F0* vs ln *CP* (log-log scale).

These equations clearly show that *OQ* increases with an increase in pitch while *SQ* decreases in value when pitch increases.

The logarithmic transformation was used in Eqns. 7.2 and 7.3 because it produces a better best-fit curve than a direct linear equation. Figure 7.7(a) shows the linear best fit for a scatter plot of *SQ* vs *F0* for subject 1. Figure 7.7(b) shows the same plot with a log-log best-fit line. It is clear that the log-log line is a better fit than the linear plot.

Relationship between *OP*, *CP*, *sCP* and *F0*

The shape of the EGG waveform can be parameterised in terms of *OP*, *CP* and *sCP*. As with *OQ* and *SQ*, these parameters were also fitted with logarithmic curves. These parameters form the basis of the new model presented in section 5.3. The plots in Figure 7.8 and Figure 7.9 indicate that ln *OP*, ln *CP* and *sCP* have a negative relationship with ln *F0* for all subjects studied. The relationship between these parameters and *F0* for both male and female subjects by linear regression are as shown in table 7.3.

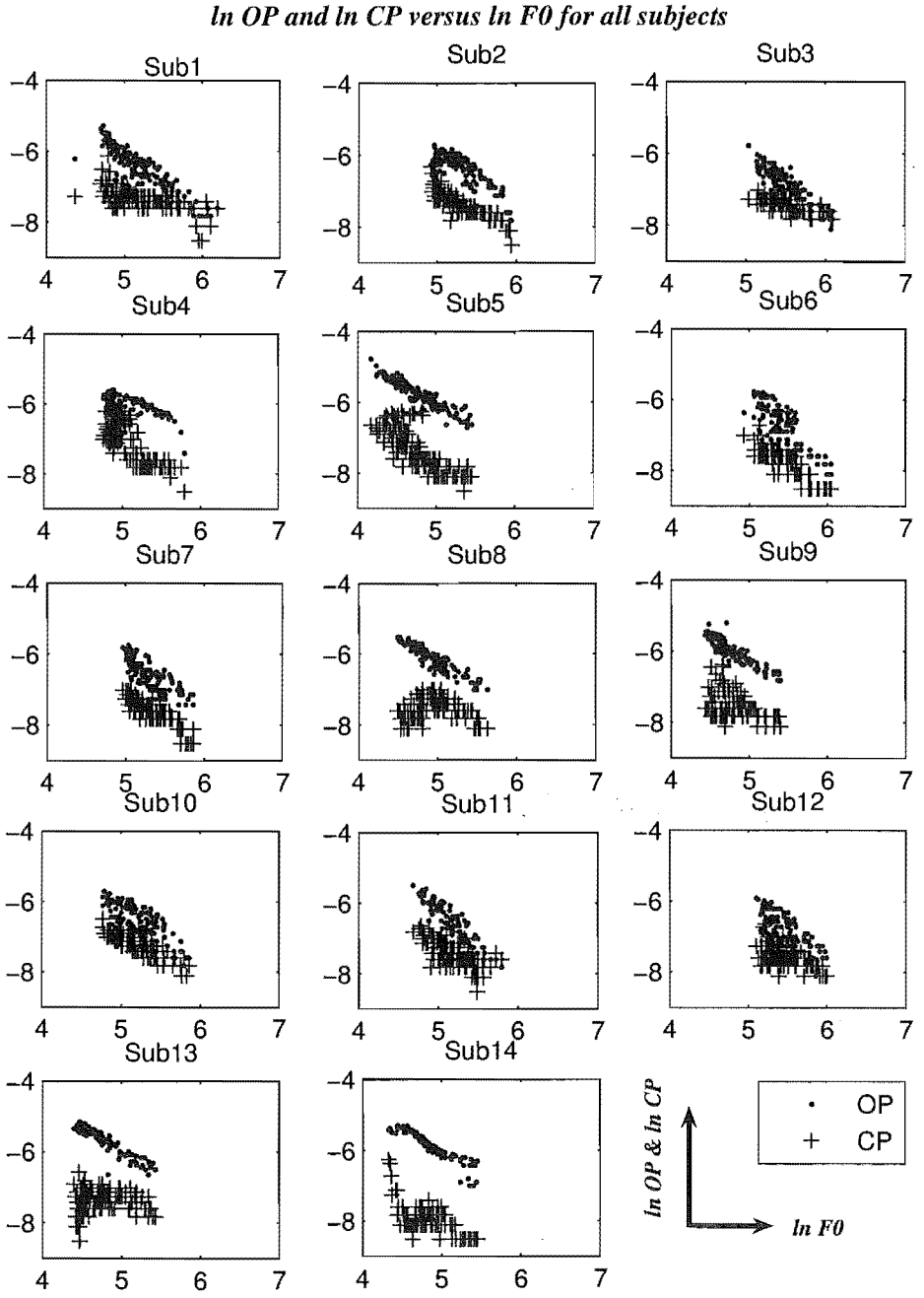

Figure 7.8 Opening phase and closing phase versus ln *F0* plots for all subjects.

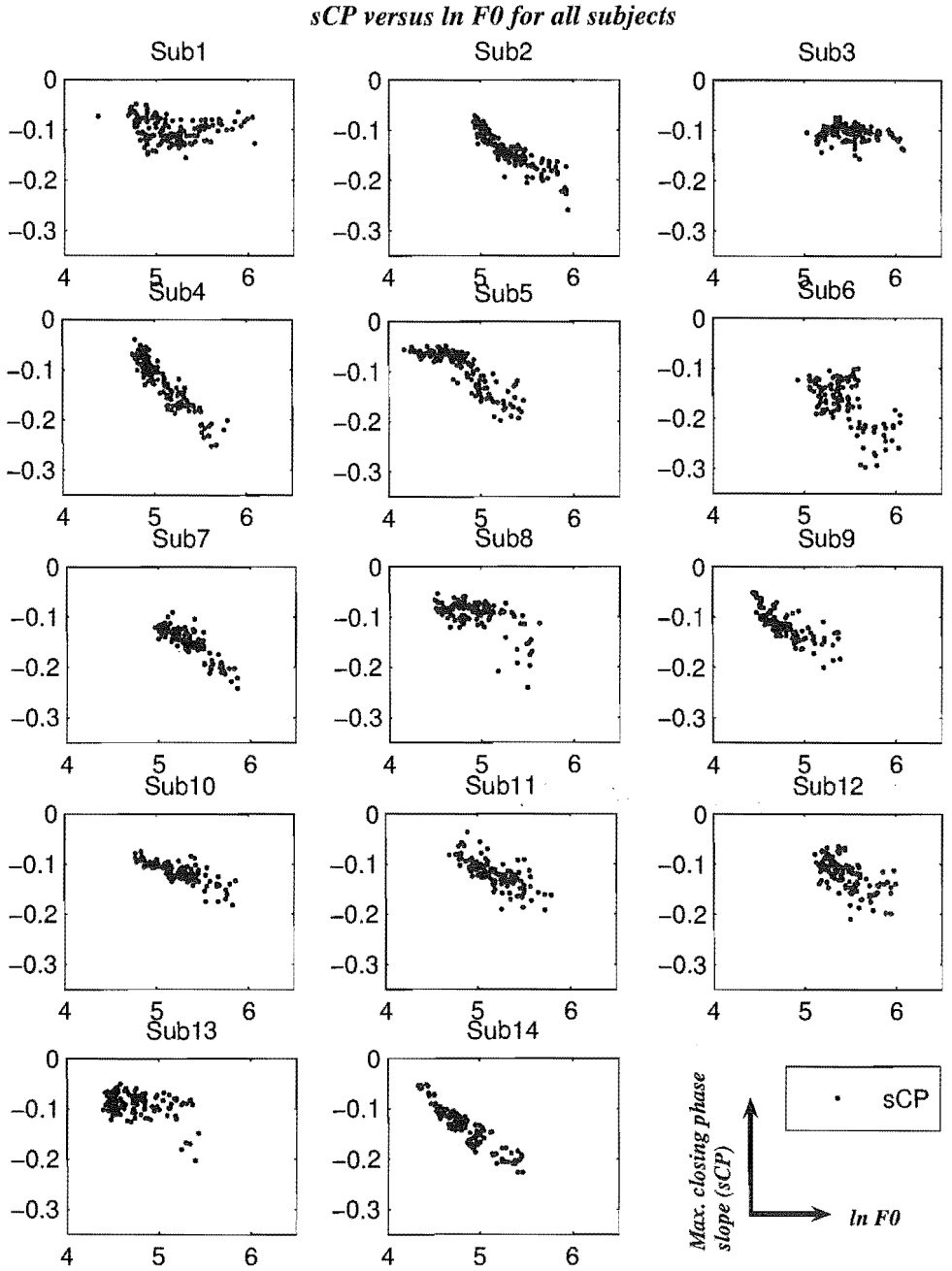

Figure 7.9 Maximum slope of the closing phase versus ln *F0* plots for all subjects.

Table 7.3 Linear regression equations of parameters for male and female subjects

	ln *OP*	ln *CP*	*sCP*
M	-1.09ϕ-0.7	-0.75ϕ-3.83	-0.08ϕ+0.276
F	-1.39ϕ+2.07	-0.93ϕ-2.51	-0.076ϕ+0.276
All	-1.24ϕ+0.69	-0.84ϕ-3.17	-0.078ϕ+0.276

Note: M = male subjects, F = female subjects and ϕ = ln *F0*

Voice registers

Almost all the recorded voices from the 14 subjects were judged by the experimenters as being modal register based on visual inspection of the waveforms and also audio appraised of the recorded acoustic signals. Double opened phase vocal fry was observed in 20 out of the 1901 ($\sim$1 %) analysed voice tracts. Contrary to what is reported in the literature, the *F0* for vocal fry was actually much higher in some of these subjects than the *F0* values reported in the literature [BCNG98, WMW84, Hol68]. Possibly those cases were not pure vocal fry, but a transition from modal to vocal fry. No falsetto mode was observed. This shows that indeed vocal fry and falsetto are seldom used in normal speech, consistent with results reported by Hollien [Hol68]. Comparison between the average pulse shape of the modal register and vocal fry reveals that all parameters in both cases decrease with increasing *F0*. However, all the parameters for vocal fry are lower than those of the modal register for *F0* below 110 Hz. Above 110 Hz, *sCP* for vocal fry is higher than modal register. It may be possible that this is the point where the voice register changes from vocal fry to model mode.

7.1.5 Conclusions

Through this experiment, the key features of the EGG signals that can be used for EGG waveform modelling were able to be extracted. All parameters: ln *OP*, ln *CP* and *sCP*, showed a negative relationship to ln *F0*. We were also able to show that the EGG signal is independent of vowel effect. Analysis of the EGG waveform for both speaking and singing range within $\pm$ 1 octave of the habitual pitch show no significant difference in the parameters measured. It was also observed that model voice is used for both speaking and singing.

The parameters used for the twin-bar model introduced in Chapter 5 are based on the results of this experiment. The novelty of this model is the use of *F0* as the only input to the design to produce glottal pulse, as all other parameters are either predefined or can be calculated by selecting the required *F0* value. The twin-bar model has a working range of 65-270 Hz for male, 195-590 Hz for female and, 120-410 Hz if both male and female data were combined. This is also the only model thus far that changes its waveform shape according to the change in *F0*.

The next section compares the twin-bar model with two other well-known models in providing the

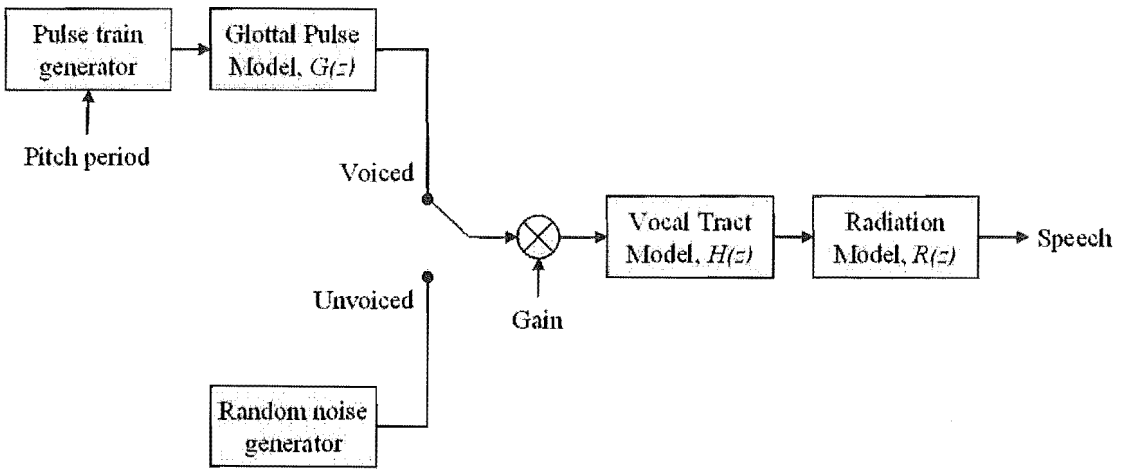

Figure 7.10 A general discrete-time model for artificial speech production.

glottal source for voice synthesis.

7.2 Voice synthesis

This section describes the voice synthesis process. Speech synthesis is necessary as it is not possible for a person to provide a fixed vocal tract shape (filter) for the production of artificial speech with three different glottal models. The half-sample delay Kelly-Lochbaum vocal tract model provides the constant vocal tract required (see section 5.4).

Figure 7.10 shows the discrete time speech production system. In voiced speech such as vowel sounds, the vocal tract, $H(z)$, and lip radiation, $R(z)$, are excited by the glottal source, $G(z)$. In unvoiced speech, the excitation source is a random noise generator. Since we are only interested in voiced sounds, only vowel sounds will be addressed here. The three glottal source models, $G(z)$, used for the vowel synthesis are the Rosenberg model, the LF model and the twin-bar model (refer to sections 5.2.2, 5.2.3 and 5.3 respectively).

The vocal tract area function used in this research is based on the work by Story *et al* [ST96] where the values of the area functions were obtained from magnetic resonance imaging (MRI) images. The length of each vocal tract segment in their study is 3.97mm. The length of each segment is the distance sound travels in half a sample period ($\frac{1}{2f_s}$):

$$L = \frac{c}{2f_s} \tag{7.4}$$

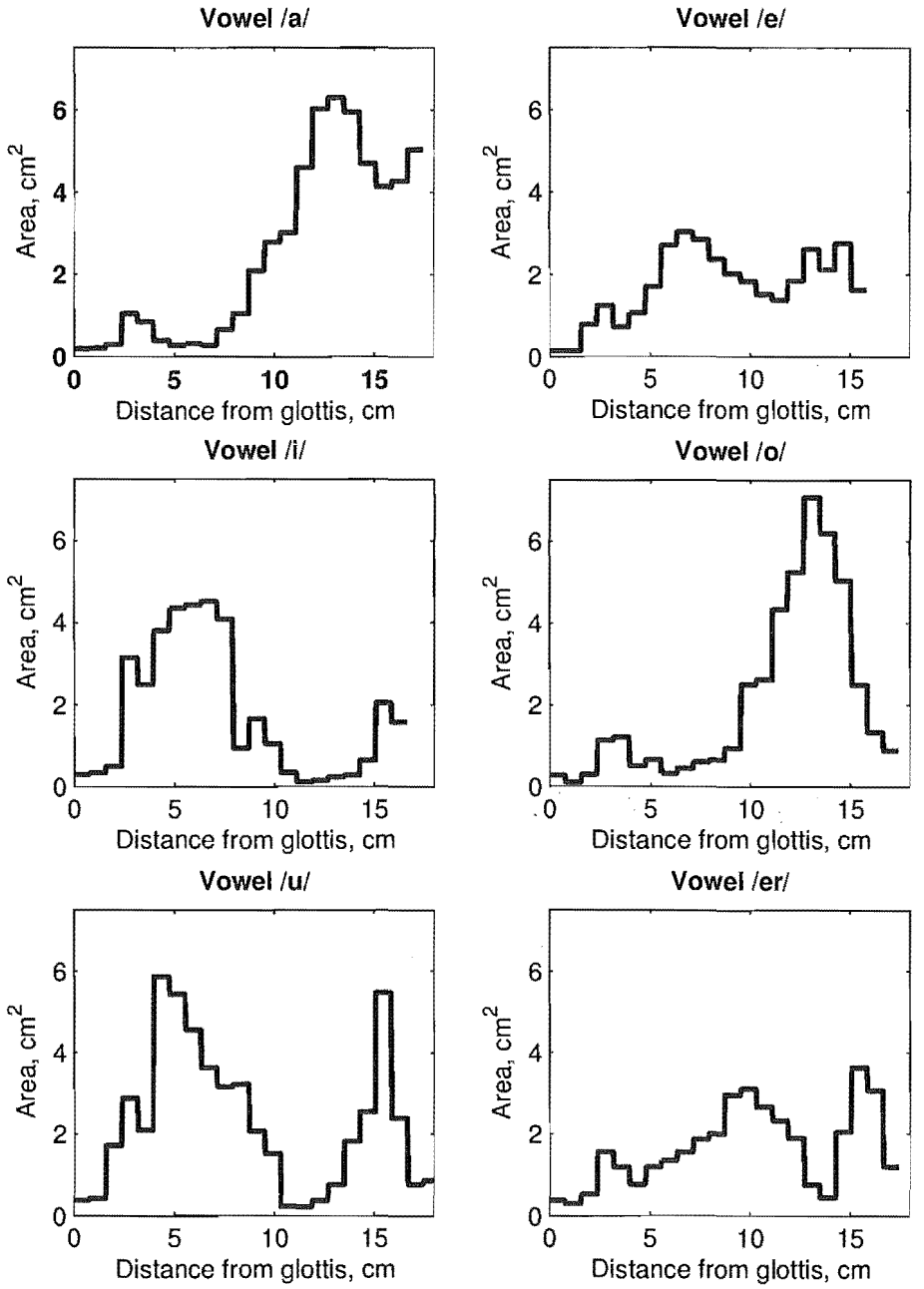

Figure 7.11 The vocal tract area functions obtained from Story *et al* [ST96] for the 6 vowels under test.

For this study, the sample rate, f_s = 22.05kHz, is half the rate used by Story *et al.* So the interval between each section becomes 7.94mm. Therefore, the vocal tract area function in this study consists of every second value of the area functions published by Story *et al.* Figure 7.11 shows the vocal tract area functions for the 6 vowels under test. Figure 7.12 shows the vocal tract transfer functions of the 6 vowels that correspond to the vocal tract area function in Figure 7.11. The formants of the vowels generated were consistent to with those reported in the literature [BF04, Mac00].

The vocal tract model used for voice synthesis is the half-sample delay Kelly-Lochbaum structure with one-multiply junction (refer to section 5.4.3). The algorithm for this model is shown in the flowchart in Figure 7.13. The glottal source, $g[n]$, is the input to the K-segment vocal tract area function of a vowel sound, $\{A_i, i = 1 \text{ to } K\}$, where each segment corresponds to a half-sample delay. *Pout* is the synthesised vowel sound at the lips. The synthesised voice is produced when this signal passes through the lip radiation filter, *R(z)* (as described in section 5.4.5).

Assumptions made for this model are:

- The vocal tract consists of a series of concatenated lossless tube sections.
- The reflection coefficients at the glottis, r_g, and at the lips, r_l, are set at 0.8 and -0.8 respectively (as discussed on section 5.4.2).
- The same glottal pulse jitter pattern can be used for all vowels. The pattern incorporated in the synthesised voice was extracted from a normal subject's EGG signal and was used to provide the *F0* variation.

7.3 Perceptual tests

Perceptual tests were carried out to compare the synthesised voice for different glottal models. Two separate tests were conducted to determine the identity and the quality of the synthesised voice using the Resenberg model, the LF-model and the twin-bar model.

7.3.1 Setup

The experiment was carried out in a quiet room. The digital synthesised voice was converted to analog signal via a Conexant AMC Audio sound card (analog frequency response of 20-20kHz) on ACER's Aspire 1680, 1.6GHz Pentium Centrino laptop. A Bassonic headphone (manufactured in China) with frequency response 18Hz-20kHz was used to provide the sample sounds for the subject to listen to.

A perceptual test software written by the author using Visual Studio C++ was used to provide a graphical user interface (GUI) for the subject and also to generate test files containing the test results.

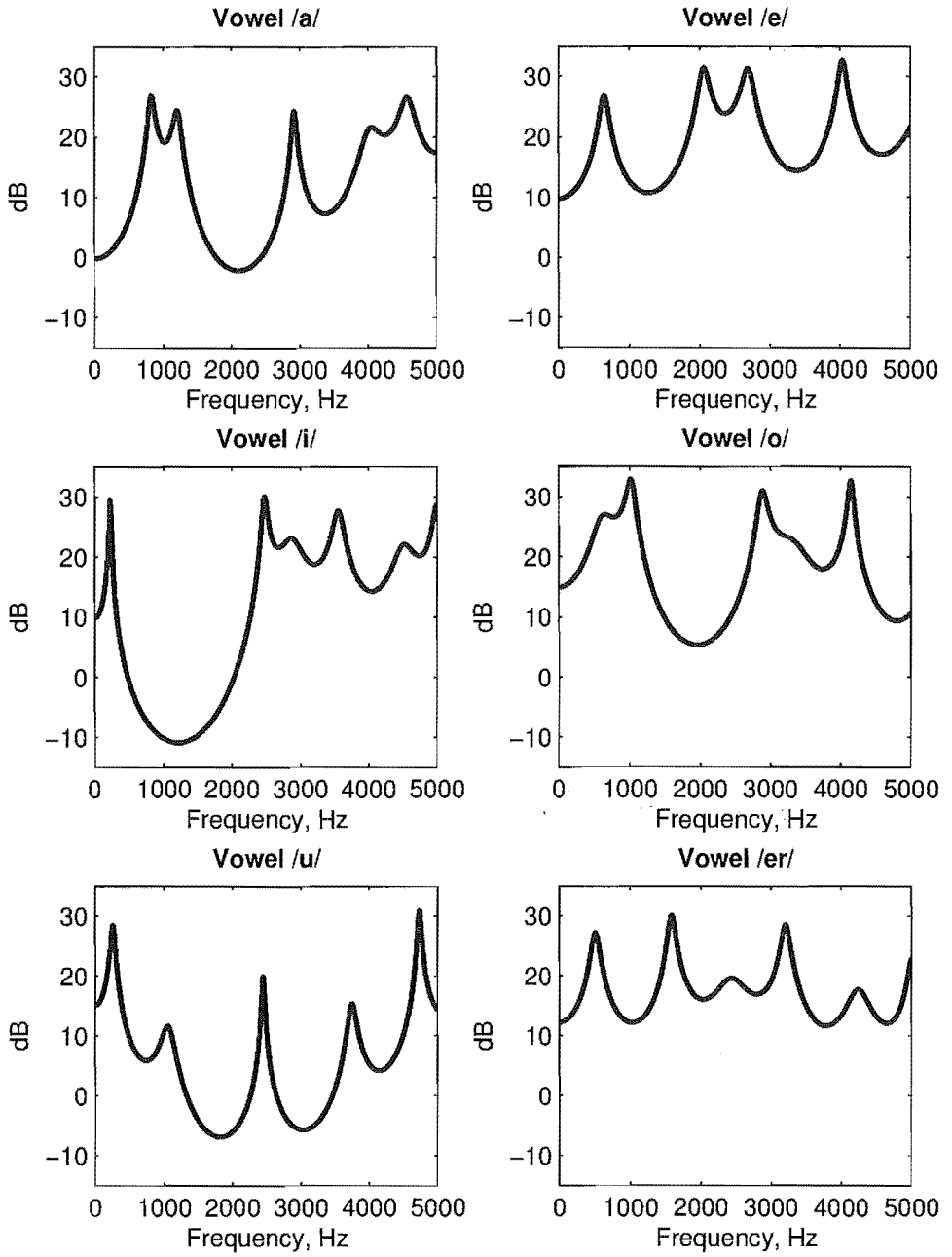

Figure 7.12 The vocal tract transfer functions for the 6 vowels.

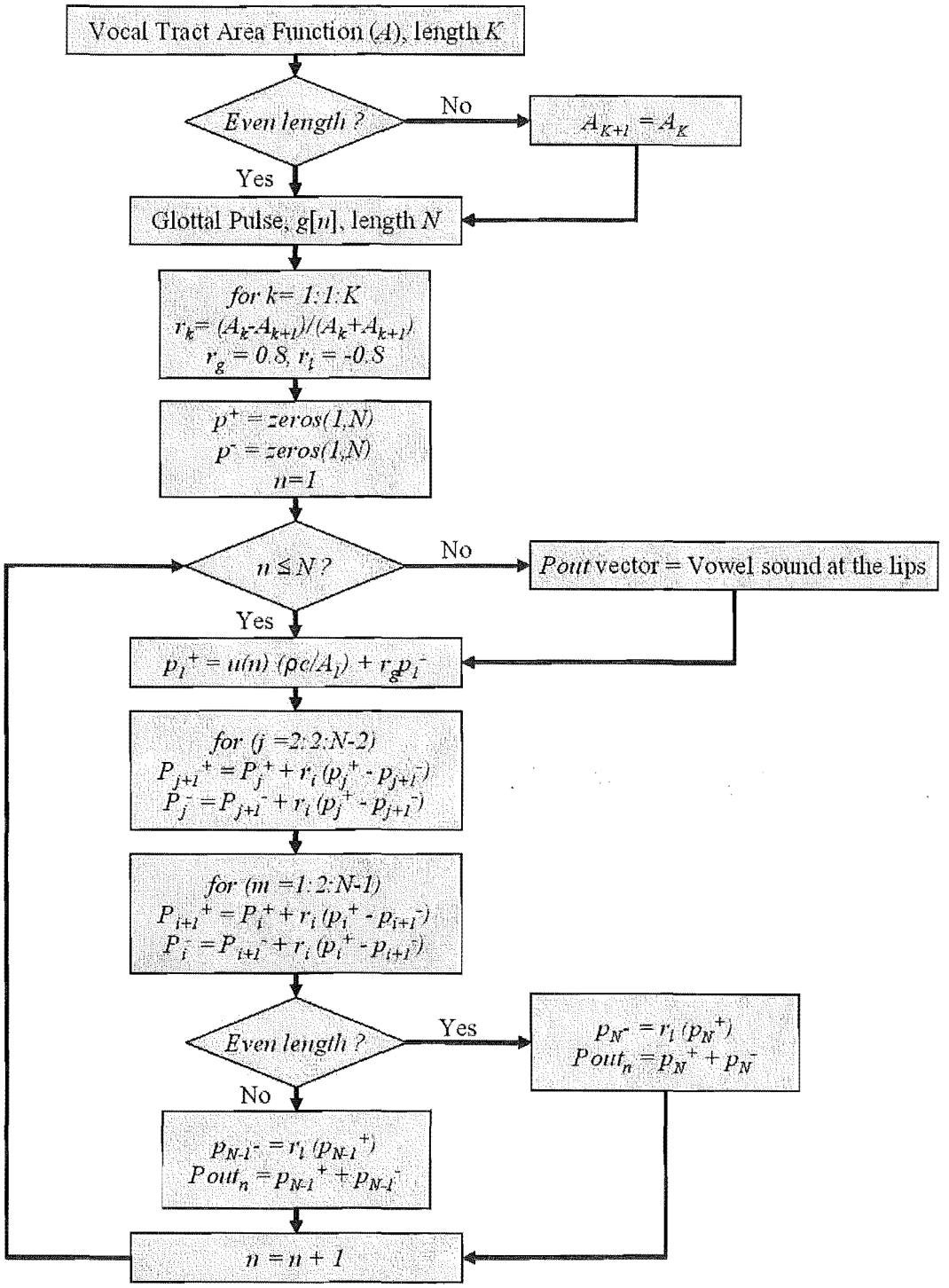

Figure 7.13 The flow chart for voice synthesis with glottal source, $g[n]$ as input to a K-segment vocal tract model (each segment is equivalent to $\frac{1}{2}$ sample delay).

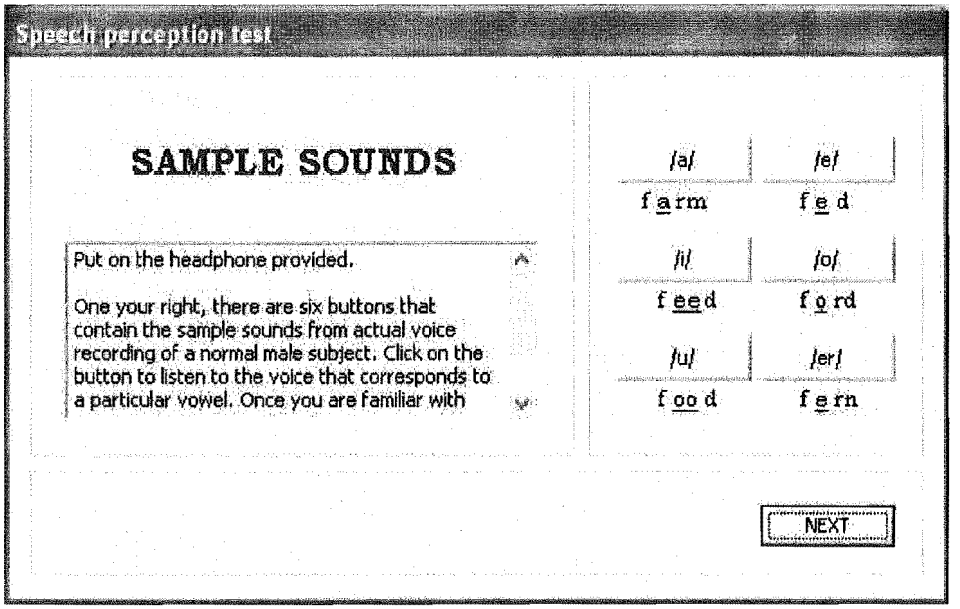

Figure 7.14 The GUI for perceptual tests, showing the vowels /a/, /e/, /i/, /o/, /u/ and /er/. Under each button is a hint showing the subject what the vowel should sound like. A subject can listen to a particular vowel sound by 'clicking' on the corresponding button.

7.3.2 Procedure

The subjects for this experiment consisted of 10 adults (5 males and 5 females) with no hearing disorders. The average age was 24.4 years (SD 3.10 years). Subjects were recruited through advertisement within the University of Canterbury.

Before the tests were carried out, the subjects were allowed to familiarise themselves with the vowel sounds by clicking on the buttons in the GUI (see Figure 7.14) that correspond to the six sample sounds from actual voice recordings of a normal male subject. Below each button is a word that provides a hint for the pronunciation of the vowel symbols.

Once the subject was confident that they could recognise the sounds, they proceeded to the first test: the vowel identity test. In this test, the subject was asked to listen to 18 wave files (6 vowels $\times$ 3 glottal models) through a headphone and to decide which of the 6 vowel sounds they belonged to by clicking the corresponding button on the screen. The hints below the vowels were provided to give the subject an idea of what the vowels should sound like. The subject could listen to the given sound more than once by clicking the 'Replay' button. If the subject could not decide which vowel a sound belonged to, he/she clicked on the 'Cannot Decide' button. If the subject accidentally clicked on the wrong button, they could click the 'Back' button, which allowed the subject to listen to the previous sound. The test was carried out 5 times with a total of 90 sample sounds (6 vowels $\times$ 3 glottal models $\times$ 5 trials).

The second test was the vowel quality test: given the target vowel, the subject was required to decide which of the three synthesised vowel sounds was more similar to a natural male voice. The subject was given two sounds at a time to compare. The two sounds were randomly chosen and played through the headphones. The subject was required to decide which of the two synthesised vowel sounds was closer to a natural male voice by clicking on the button 'A' or 'B'. The sounds could be replayed by clicking the 'Replay A' or 'Replay B' button. If the subject could not decide which of the two sounds was better, he/she clicked on the 'Cannot Decide' button. If the subject accidentally clicked on the wrong button, they could go back and listen to the previous two sounds by clicking the 'Back' button. There were a total of 72 samples in this test (6 vowels $\times$ 6 sequences $\times$ 2 trials).

In each test, synthesised voice generated using all three models were randomly chosen and played through a headphone. The six vowels tested in this experiment include: /a/, /e/, /i/, /o/, /u/ and /er/. The overall duration for this experiment was approximately 20-25mins.

7.3.3 Statistical analysis

Data from each subject in the first test was categorised into the number of correctly identified vowels (T), incorrectly identified vowels (F) and undecided (U) for each glottal model and each of the 6 vowels (see Table 7.4). Before the data was analysed, they were converted into percentage (see Table 7.5) using the following equation:

$$Data(\%) = \frac{number\ of\ \mathrm{T}\ or\ \mathrm{F}}{total\ number\ of\ trials\ -\ number\ of\ \mathrm{U}} \times 100 \tag{7.5}$$

where *total number of trials* = 5.

Table 7.4 Statistical analysis - True/False vs Glottal Model

Subject 2	Twin-bar			LF			Rosenberg		
	T	F	U	T	F	U	T	F	U
/a/	5	0	0	5	0	0	5	0	0
/e/	5	0	0	5	0	0	5	0	0
/i/	5	0	0	3	1	1	5	0	0
/o/	3	1	1	5	0	0	4	0	1
/u/	5	0	0	4	0	1	2	1	2
/er/	5	0	0	5	0	0	5	0	0

All data were submitted to a series of chi-square tests to determine if the ability of a person to identify a set of vowels and the glottal model used to generate them were related or independent of each other.

Table 7.5 Statistical analysis - True/False vs Glottal Model

Subject 2	Twin-bar		LF		Rosenberg	
	T (%)	F (%)	T (%)	F (%)	T (%)	F (%)
/a/	100	0	100	0	100	0
/e/	100	0	100	0	100	0
/i/	100	0	75	25	100	0
/o/	75	25	100	0	100	0
/u/	100	0	100	0	67	33
/er/	100	0	100	0	100	0

A similar procedure was carried out for vowel quality test to determine whether vowel quality and the glottal pulse model used to generate the vowels were related or independent. The data for this test is categorised into the number of "more natural" (M), "less natural" (L) and " cannot decide" (U) sounds, for each glottal model, given the identity of the vowel. As with the vowel identity test, they were also converted to percentage before data analysis begins:

$$Data(\%) = \frac{number\ of\ \mathrm{M}\ or\ \mathrm{L}}{total\ number\ of\ comparisons\ -\ number\ of\ \mathrm{U}} \times 100 \tag{7.6}$$

where *total number of comparisons = 48.*

7.3.4 Results and discussions

The results of the statistical analysis for the vowel identity and vowel quality tests are presented separately.

Vowel identity test

Figure 7.15 shows the percentage of vowels that were correctly identified by each subject for each of the 3 glottal models.

The chi-square test on combined data from all subjects shows a chi-square value of 1.609 (P=0.447), which implied that there was no relationship between vowel identity and the type of glottal pulse model that is used to generate them. In other words, none of the glottal models were superior, in terms of synthesising vowel sounds that can be more easily identified. This is possibly because the identity of a vowel sound is largely dependent on the vocal tract area function (e.g the first three formants of the spectrum), not so much on the glottal source. Since the vocal tract shape was fixed for a particular vowel, the formants for that same vowel for all three models were the same.

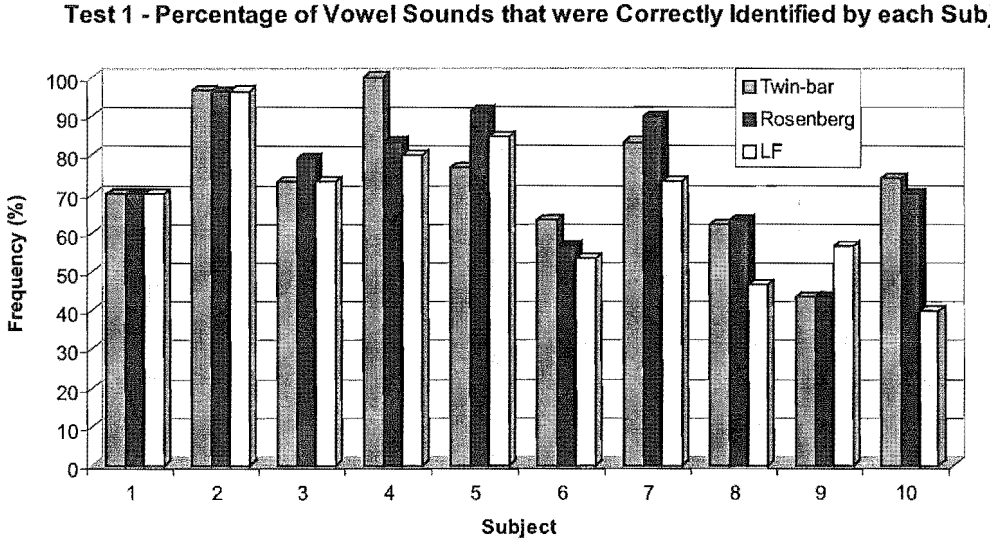

Figure 7.15 The percentage of vowel sounds that were correctly identified by each subject for the twin-bar model, Rosenberg's model and the LF model.

Subject dependent: Analysing the data on the subject level reveals that 50% of the subjects showed that the two variables were related (refer to Table 7.6) and the other 50% showed the two variables were not related. Of the ones that showed significant relationship between vowel identity and glottal model for separate subjects, three out of five subjects revealed that vowel sounds generated with the Rosenberg model were more easily identified than the other two models; two out of five showed that the twin-bar model was better.

Vowel effect: To test whether there is a vowel effect on the results, the data analysis was also carried out for separate vowels (Table 7.7). Four out of the six vowels tested (/i/, /o/, /u/ and /er/) showed that vowel identity were significantly related to the glottal models used to generate them; vowels /i/ and /o/ were found to be easier to identify using the Rosenberg's glottal model while the twin-bar glottal model was found to be better for identifying vowels /u/ and /er/. The other two vowels (/a/ and /e/) showed no vowel effect on the results.

Vowel quality test

Figure 7.16 shows the percentage of vowel sounds that were perceived to be better by each subject when vowel sounds generated by the 3 glottal models were compared.

The chi-square test on the combined data from all subjects for the vowel quality test shows a chi-square value of 38 ($P<0.001$), which implies that vowel quality and the glottal models are signif-

Table 7.6 The chi-square test results for vowel identity versus glottal models for individual subjects.

Subject	Chi-square	P
1	0.00	1.000
2	0.19	0.910
3	1.28	0.527
4	21.52	<0.001
5	7.47	0.024
6	2.08	0.354
7	9.89	0.007
8	6.57	0.037
9	5.24	0.073
10	29.12	<0.001

Table 7.7 The chi-square test results for vowel identity versus glottal models for individual vowels.

Subject	Chi-square	P
/a/	2.03	0.363
/e/	4.92	0.085
/i/	52.45	<0.001
/o/	7.61	0.022
/u/	38.44	<0.001
/er/	8.45	0.015

icantly related. The twin-bar model produced sound quality that were significantly better than the other two glottal models, presumably because this model took into account the changes in waveform shape as pitch is varied. It was followed by the Rosenberg's glottal model and then the LF glottal model.

Subject dependent: The chi-square test of vowel quality versus glottal models for individual subjects, the vowel quality and the glottal models were significantly related for all subjects, except for subject 7. Of the nine subjects, eight of them showed that the twin-bar glottal model was better. The other one showed that both the twin-bar model and the LF model were equally good.

Vowel effect: The chi-square test of vowel quality versus glottal models for individual vowels revealed that the two variables were significantly related for all vowels. Five vowels (/a/, /e/, /o/, /u/ and /er/) showed that the vowel quality generated with the twin-bar model produced more natural sounding voice than the other two models. Vowel /i/ showed that the Rosenberg model was better.

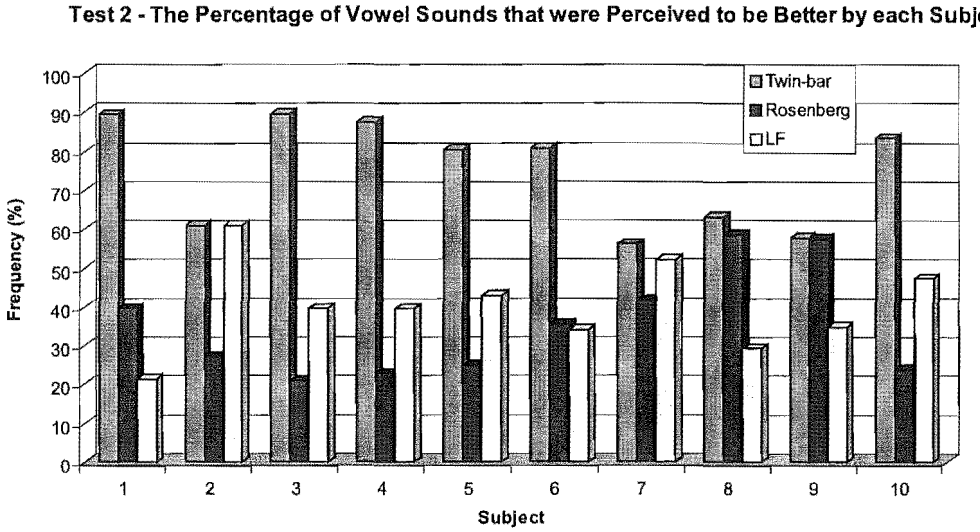

Figure 7.16 The percentage of vowel sounds that were judged to be better by each subject when vowel sounds generated by the twin-bar model, Rosenberg's model and LF model were compared.

7.3.5 Conclusion

The first part of this study suggests that the identity of a synthesised vowel sound is not significantly related to the three glottal source models used to generate the vowel sounds; none of the models are significantly better than the other in terms of vowel identity.

The vowel quality test on the other hand showed that the three glottal source models produce sounds with different sound qualities. Of the three models, the twin-bar model was found to generate the most natural sounding artificial voice. This suggests that the twin-bar model is a good option for the sound source model of an artificial larynx to improve the quality of the artificial speech.

Chapter 8

Hardware development for MyVoice, the artificial voice device

MyVoice is the name given to the artificial speech device developed as part of the research for this thesis. *MyVoice* is designed to work inside the vocal tract: in the pharynx or in the mouth cavity. Depending on the location of the sound source in relation to the vocal folds, the missing sections can be modelled together with the glottal pulse. The vocal tract model used for the artificial voice simulation in Chapter 7 is replaced by the user's actual vocal tract. This chapter describes two prototype designs (*MyVoice1* and *MyVoice2*), including preliminary studies to test the voice quality of these designs.

8.1 *MyVoice1*: the first prototype

The first prototype was designed as a proof of concept (see Figure 8.1 for the *MyVoice1* prototype). *MyVoice1* has a built-in variable frequency controller and volume controls (one for the source signals and the other for the volume of the speech output).

8.1.1 Design layout

The *MyVoice1* consists of 4 main sections: the power supply, the pitch controller module, main circuit, and the microphone and amplifier circuits (see Figure 8.2).

The power supply

The power supply for *MyVoice1* is a 9V battery. It is converted into 3 different levels for different purposes:

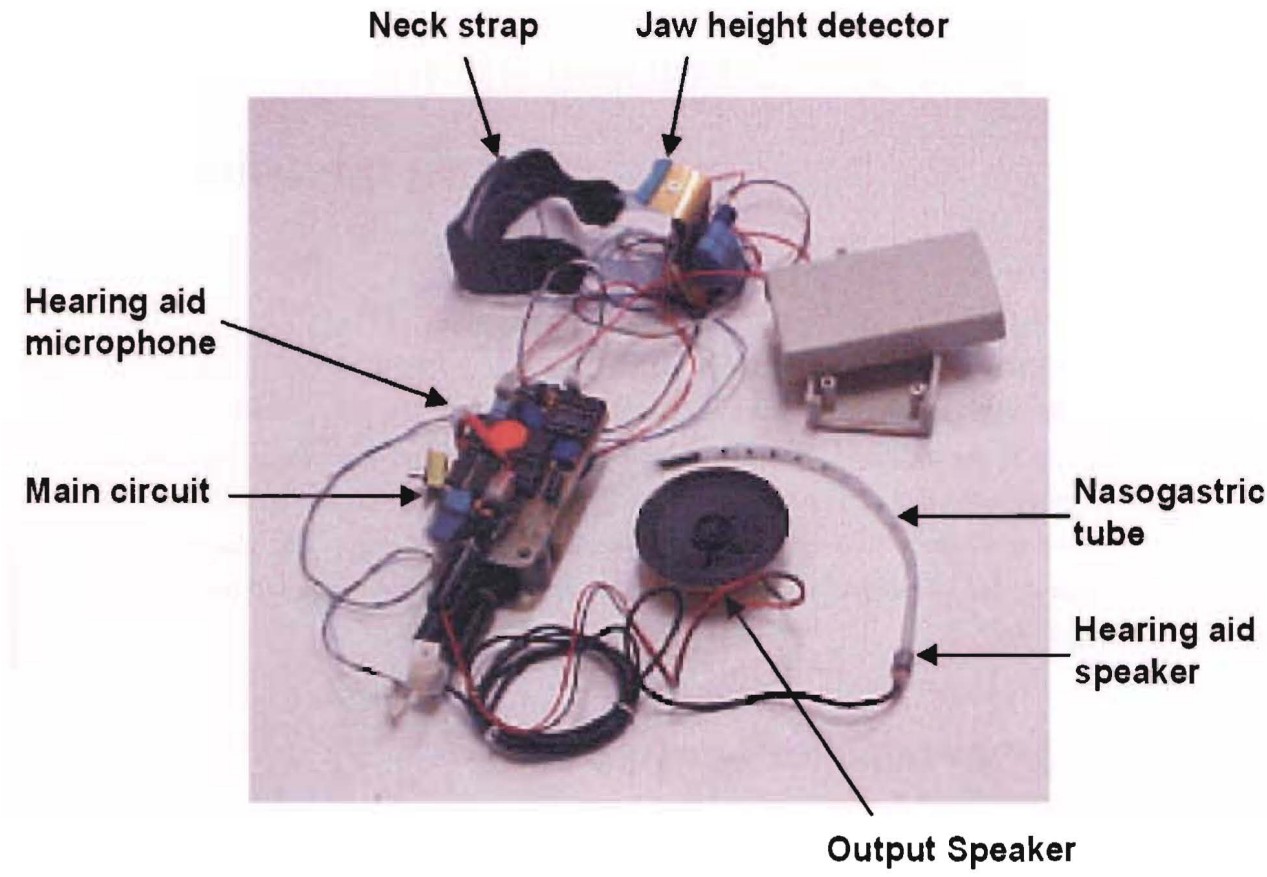

Figure 8.1 *MyVoice1* device.

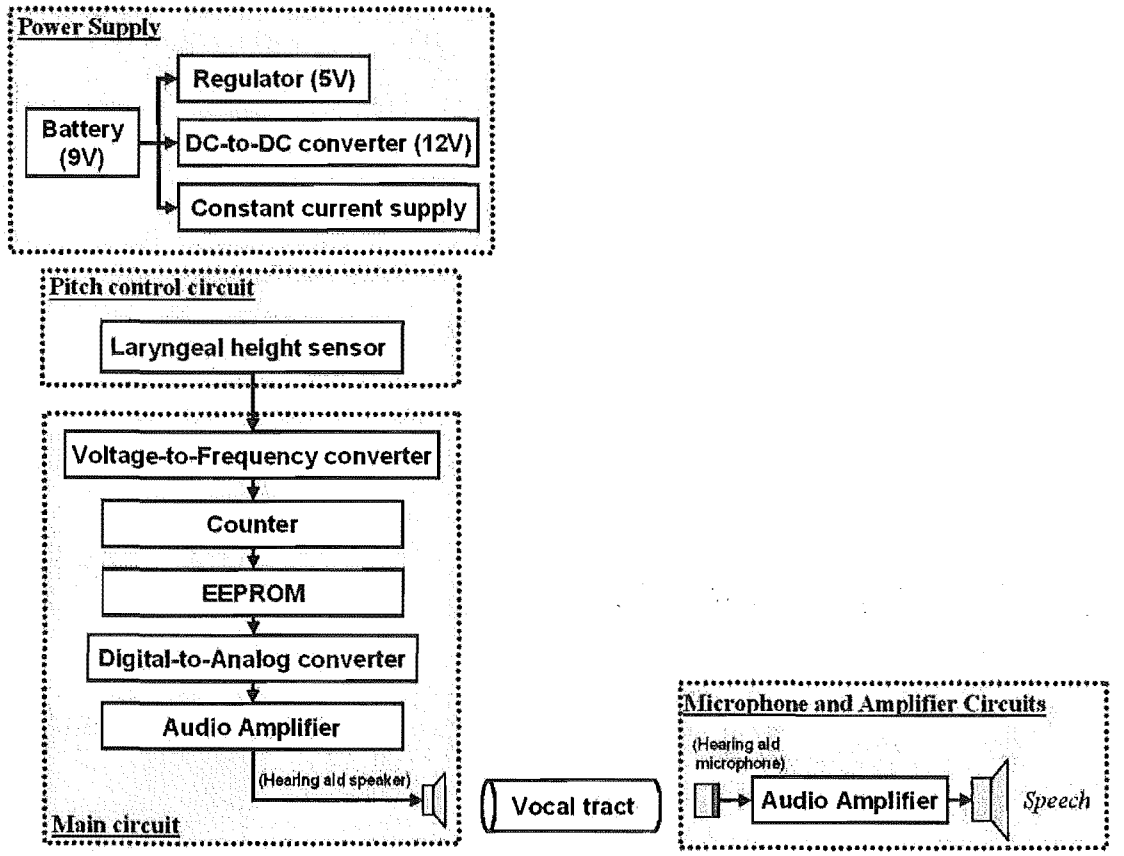

Figure 8.2 *MyVoice1* design layout.

- The 9V power supply from a battery is converted to 5V by a regulator (78L05) and is used as the main power source for the prototype.
- The 5V regulated supply is converted to 12V using a dc-to-dc voltage converter. This 12V dc supply is used for the audio amplifiers (LM386).
- A constant current supply is generated using a LM317L for the hall-effect sensor (HES) on the pitch controller.

The pitch controller module

The pitch controller module in *MyVoice1* is a laryngeal height detector made from a brass lever, a magnet and a HES. The module was designed such that it can be fitted onto the connector on a tracheostomy tube. When in use, the pitch controller is placed just below the larynx. The laryngeal height detector is used to control the clock speed of the pulse generator, which effectively controls the pitch of the glottal pulse. The frequency range is between 100 and 400 Hz.

The main circuit

The output of the HES is sent to a voltage-to-frequency converter (VFC, XR4151) and then a counter (HC4024). The output of the counter is connected to the input of the EEPROM (NMC27C16) where an 8-bit averaged EGG waveform is stored as glottal pulse template. Data from the EEPROM is sent to a digital-to-analog converter (DAC, DAC0832LCN) at each increment of the counter. When the counter reaches 100 (end of the glottal pulse), the counter is reset and the cycle begins again. The analog signal at the DAC is amplified (using LM386) and sent to a hearing aid speaker. The amplified signal then goes to a speaker which is in turn connected to one end of a nasogastric tube. The sound that comes out of the other end of the nasogastric tube is the sound source of *MyVoice1*.

Microphone and amplifier circuits

As the user "mouths-the-words", the microphone placed just outside the lips picks up the voice. This signal is then amplified before being sent to a larger speaker where it can be perceived as speech.

8.1.2 Prototype testing

A test using the prototype was carried out on a normal subject. The pitch controller, attached to the tracheostomy tube (the end of the tube that goes into the trachea is cutoff, see Figure 8.3) was strapped on the subject's neck making sure that the lever was placed just below the larynx. The free end of the nasogastric tube was placed inside the mouth cavity. The subject mouthed the words: "Hello. How are you"? Results of the test was observed by 5 professionals (one medical doctor, two speech therapists and two biomedical engineers) who agreed that although the artificial speech produced was perceivable, *MyVoice1* suffered from a few obvious drawbacks:

Figure 8.3 An enlarged version of the laryngeal height sensor used in *MyVoice1*. The tracheostomy tube going into the trachea has been cut off so that a normal subject can use the laryngeal height sensor for testing.

- Even on a normal subject the larynx did not move very much when "mouthing the words". Therefore the required pitch variation could not be reliably obtained.
- The pitch range was not limited. Normal speaking range is ± 4 semitones from habitual pitch. The pitch range for *MyVoice1* varies anywhere from 80Hz to 400Hz, depending on the movement of the larynx.
- The pitch variation was not based on experimental results. As a consequence to that the artificial speech generated still did not sound natural.
- ICU patients who have tracheostomy tubes in their throat are unable to move their larynx even when they try because the tracheostomy tube limits the desired movement.

On the basis of what was learned from preliminary tests with *MyVoice1*, a number of experiments were carried out to find the most appropriate method of controlling pitch when "mouthing the words" (Chapter 6) and a glottal sound source that is a better approximation of the function of the vocal folds (Chapters 5 and 7).

8.2 *MyVoice2*: the second prototype

MyVoice2 is an improved version of *MyVoice1*. Two major improvements on the original design are: (i) the incorporation of the twin-bar model, a glottal pulse model that allows the glottal pulse shape to vary with pitch, and (ii) the pitch tracker/controller uses a reflective object sensor for non-contact jaw height measurement. There are also a few extra features in the new design, such as an

automatic on/off control switch, habitual pitch selection, the option to change the location of the sound source (in the mouth cavity or in the pharynx) and more sophisticated interfaces between component sections. Combined, the improvements have made *MyVoice2* a much more practical system than its predecessor.

8.2.1 The design layout

MyVoice2 is divided into 3 main sections: pitch controller, glottal pulse generator and speech enhancer, and glottal pulse transmitter and voice receiver. The block diagram for *MyVoice2* showing the separate sections is depicted in Figure 8.4.

The pitch controller

The pitch controller consists of a jaw detector sensor and a micro-controller. The jaw detector uses reflective object sensor (OPTEK, OPB706): a non-contact sensor to measure the jaw height as the user "mouths the words" (see Figure 8.5). The reflective object sensor consists of an NPN silicon photo-transistor and an IR emitting diode. It has a working range of approximately 40mm (which is within the average maximum jaw height of a speaker during normal conversation, approximately 20mm). The output of the sensor is not linear with respect to jaw height, so it has to be calibrated. The jaw height sensor calibration is performed by placing a beige coloured plastic material over the reflective object sensor. The material is placed between 0mm and 82mm from the sensor, at 1mm intervals. The ADC output at each position is measured. A lookup table that consists of the raw data and the jaw height is then generated and stored inside the micro-controller. The circuit for the pitch detector is quite straightforward (see Figure 8.6 for the schematics of the pitch controller).

The micro-controller used in this prototype is the ATmega8AI series from ATMEL. The micro-controller is used to convert the analogue signal from the jaw detector into an 8-bit digital signal. The jaw height signal is sampled at 300 times per second. Next, the lookup table stored inside the micro-controller is used to find the pre-calibrated jaw height signal before sending it to the laptop (Pentium 3, Acer, Taiwan) via the serial interface (ComPort1, baud rate 38400). Most of the signal processing is carried out by the laptop computer.

Glottal pulse generator

Although the glottal pulse generator has default values, an initial setup by the user or caregiver that matches the user's characteristics improves the quality of the sound produced. These include:

1. Gender - males and females have different glottal waveform shapes even at the same *F0* and this is taken care of by the separate equations for the glottal parameters shown in Table 5.1 and Table 5.2.

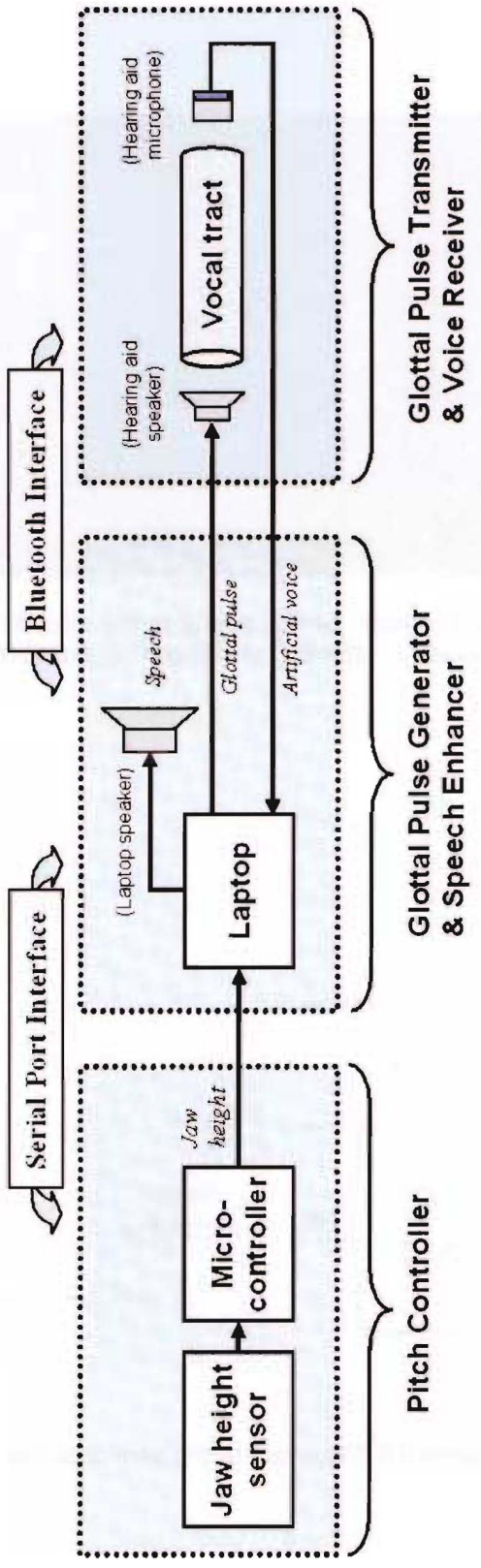

Figure 8.4 *MyVoice2* design layout.

Figure 8.5 The reflective object sensor used for measuring jaw movement in *MyVoice2*. The reflective object sensor is placed inside a hollowed out microphone casing from a speaker/microphone headset.

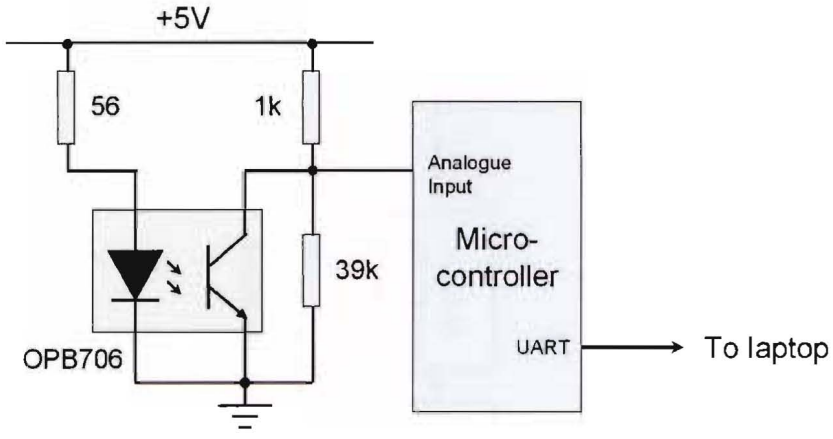

Figure 8.6 The circuit for the pitch controller.

2. Habitual pitch - different people speak at different habitual pitch level. Adjusting the habitual pitch to the subject's habitual pitch will make the artificial voice sound more like the subject's actual voice.

3. Location of the sound source - can be in the pharynx or in the mouth cavity. The glottal waveform shape is different at different location because the section of vocal tract between the glottis and the voice source has to be taken into account.

4. Jaw height calibration - as the range of jaw opening for each person can vary, jaw height calibration makes sure that a particular jaw height is mapped to a particular pitch for a given habitual pitch.

When the headset is placed on the user, the jaw detector should be placed approximately 20mm from the point of the chin (mandibular symphysis). The subject is required to 'speak' a few test sentences in order for the computer to find the maximum and minimum jaw height. The glottal pulse generator (laptop) then uses the jaw height signal and converts it to *F0* based on the information obtained from literature where during normal speech, the *F0* varies by $\pm$ 3.4 semitones from the habitual pitch [JL86, TE93, BO00]. The *F0* is then converted to glottal pulse using the twin-bar model (Chapter 5).

The onset and termination of voicing is triggered by the input from the jaw detector. Assuming that the glottal source is off, if the jaw height signal exceeds a predefined threshold (e.g. 10 ADC units), the sound source is switched on. If the jaw height signal remains constant for a period of time (0.5 second or 150 samples), the sound source is automatically switched off.

Glottal pulse transmitter and voice receiver

The glottal pulse from the laptop is sent to the speaker on the bluetooth headset (SG212 Bobo bluetooth headset from Taiwan) via the bluetooth interface. As the computer is not bluetooth enabled, a Bluetooth Universal Serial Bus (USB) dongle (Bluetooth v1.2, made in Taiwan, operation range up to 80m and data rate up to 723kbps) is used on the computer end to allow for the communication with the headset.

The speaker used on the Bluetooth headset is a hearing aid speaker from Techtronic (receiver 26A03, Netherland). This speaker is placed inside a feeding tube that goes from the nasal cavity to the back of the soft palate (in the pharynx application) or a shorter intraoral tube if it is for intraoral (mouth cavity) application. Since the speaker output on the bluetooth headset is designed for an earphone, the signal power is too weak to create resonance in the vocal cavity. A 700-mW mono low-voltage audio power amplifier from Texas Instruments (TPA721) is added before the speaker to increase the signal power.

The receiver/microphone on the bluetooth headset is used to detect the voice generated when the user "mouths the words". This signal is sent back to the computer via the Bluetooth interface where it is subjected to further processing (e.g. speech enhancement and volume control).

8.2.2 Speech enhancement

In the computer, the voice from the microphone passes through a 20th-order direct-form FIR (low pass) filter with a cutoff frequency of 4kHz. The coefficients of the filter were calculated with the help of the FDATOOL, a toolbox in MATLAB. From the output of the filter, the signal is amplified before it is sent to the computer's speaker where it is perceived as speech produced by the subject. The speech enhancement stage provides a path for more complicated speech enhancement procedures in the future (e.g. to artificially add unvoiced sounds to the voice generated before sending it out as speech).

Currently, the voice produced with speech enhancement has to be recorded, as duplex communication on the laptop (where the microphone picks up the artificial voice and the laptop speaker produces the enhanced artificial speech simultaneously) introduces echo into the system. However, realtime operation of the device is possible by increasing the volume of the hearing aid speaker where the output is used as artificial voice (bypassing the microphone and the speech enhancement stage).

8.3 Use of *MyVoice2*

This section describes in detail the speech production process for *MyVoice2* starting from the user end to the hardware and software involved (See Figure 8.7).

A picture of the prototype, *MyVoice2* is shown in Figure 8.8. From the user's point of view, it consists of a custom-made headset, a laptop, a Bluetooth (BT) USB dongle and pitch detector (PD) serial port dongle. The +5V dc power supply for the PD serial port dongle is obtained from the USB on the laptop. To operate the device, the user simply follows the following instruction:

1. Switch ON the laptop.
2. Connect the serial port dongle onto the serial port.
3. Connect the BT USB dongle onto the USB port.
4. Establish BT connection for the laptop and the BT headset (by holding the ON button on the BT headset for 8 seconds and a further 3 seconds to allow for pairing of the device with the host (laptop). Once BT connection is established, there will be a tune on the speaker. Push the ON button once to complete the connection process.

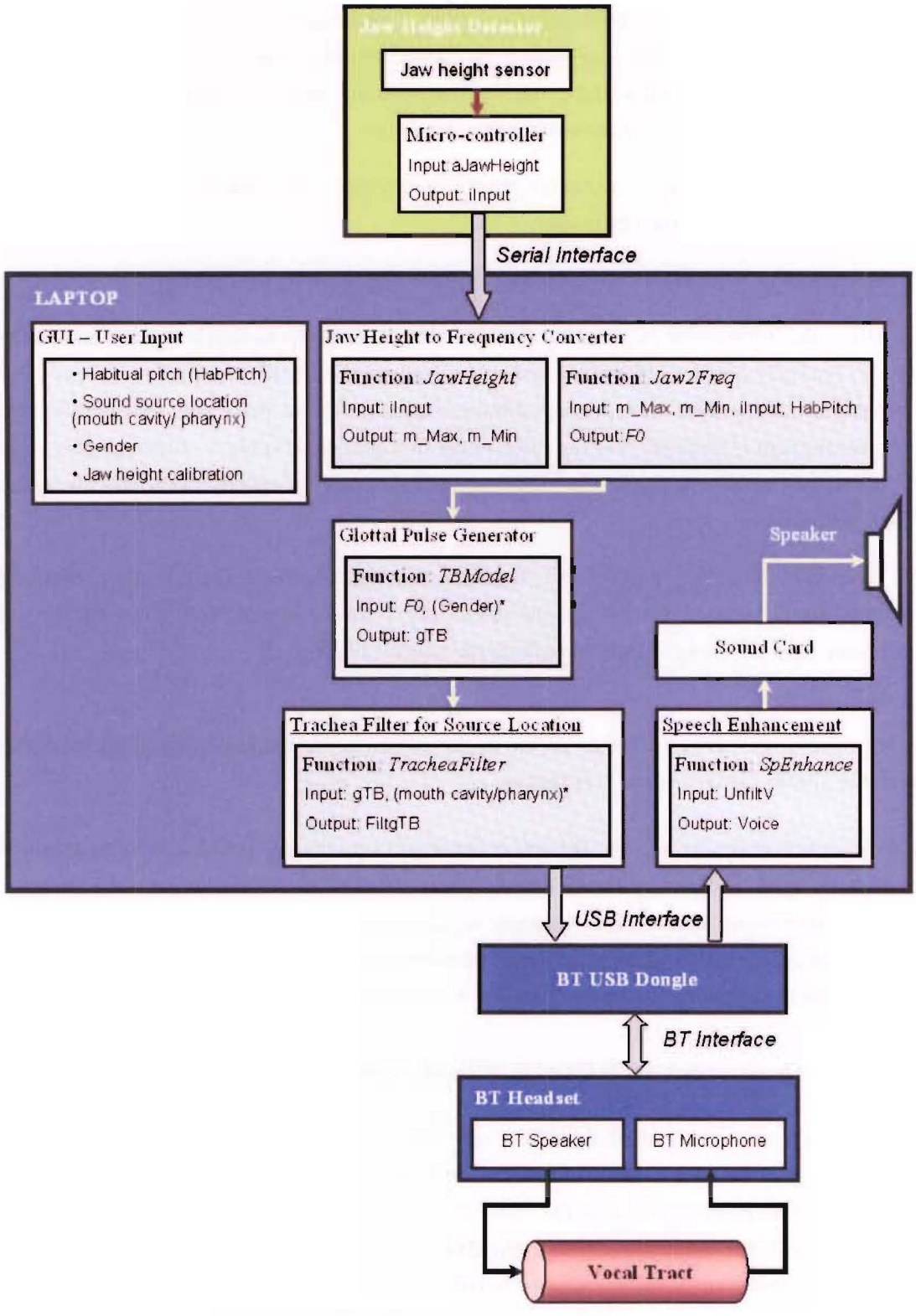

Figure 8.7 Detailed description of the speech production process for *MyVoice2*.

5. With the BT connection established, put on the custom-made *MyVoice2* headset. Place the BT speaker with its adjustable 'goose-neck' (a thin flexible structure that allows the speaker to be placed at a specific position and remains at that position) inside the mouth cavity, making sure that the speaker is pointing outwards, towards the lips.

6. Make sure the jaw height detector is approximately 20mm below the point of the chin (mandibular symphysis) when the jaw is shut.

7. Start the *MyVoice2* program by clicking the *MyVoice2* icon on the desktop.

8. Fill in the information required by the GUI. There are 3 inputs required: Habitual pitch (in Hz, typically 180Hz ±60Hz), location of sound source (inside the mouth cavity or in the pharynx) and gender. Jaw height calibration can be used to make sure that the jaw height measurement is mapped onto the pitch range of a particular subject. Although this stage is not strictly necessary, jaw height calibration will improve the quality of the speech produced significantly.

9. To operate the device, simply start "mouthing the words". Leave the jaw at a particular position for 0.5 second or more and the sound source will be turned OFF. To switch the sound source back ON, simply start "mouthing the words" again.

The information provided by the user on the GUI is stored inside the laptop and used for different parts of the glottal source production process.

When the user "mouths the words", the jaw detector picks up the jaw movement by measuring the amount of IR light produced by the IR source that is reflected from the chin onto the IR detector on the reflective object sensor. This analogue signal is converted into an 8-bit digital signal at a sample rate of 300Hz on the Atmega8AI micro-controller located inside the PD serial port dongle. A lookup table in the micro-controller is used to remove the non-linearity of the jaw height sensor. See Figure 8.9 for the jaw distance from the chin (mm) versus the raw digitised output of the jaw height sensor. The linearised signal iInput is sent to the laptop's serial port via the USART.

Inside the laptop computer a set of algorithms operate to produce the glottal sound source (see Figure 8.7). Firstly, if the user decides to calibrate the jaw height for better voice production, either the user or the care giver can click on the 'Jaw height calibration' button. The user then introduces themselves by talking/"mouthing" about themselves (e.g. by telling the other person his/her name, age, address, hobbies, etc.). As the user is "mouthing the words", the function *JawHeight* dynamically records the maximum (m_Max) and minimum (m_Min) jaw height using iInput from the serial port. The values m_Max and m_Min are stored as constants for future use. See Figure 8.10 for the flow diagram of the *JawHeight* algorithm.

Once the user has finished introducing themselves, they click on the 'Stop calibration' button. They

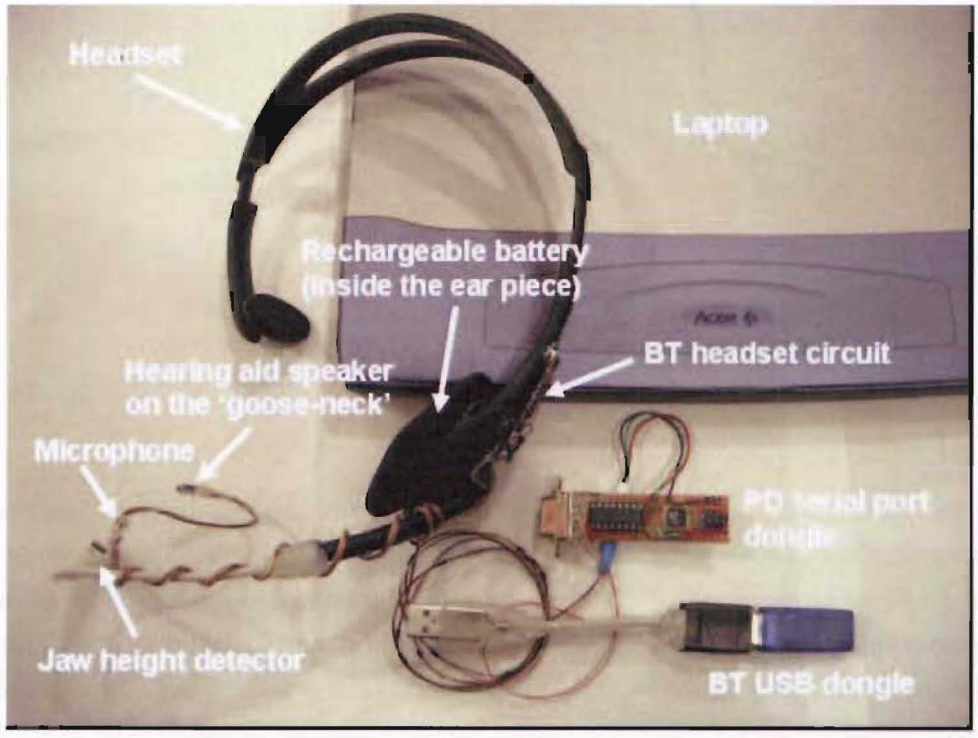

Figure 8.8 *MyVoice2.*

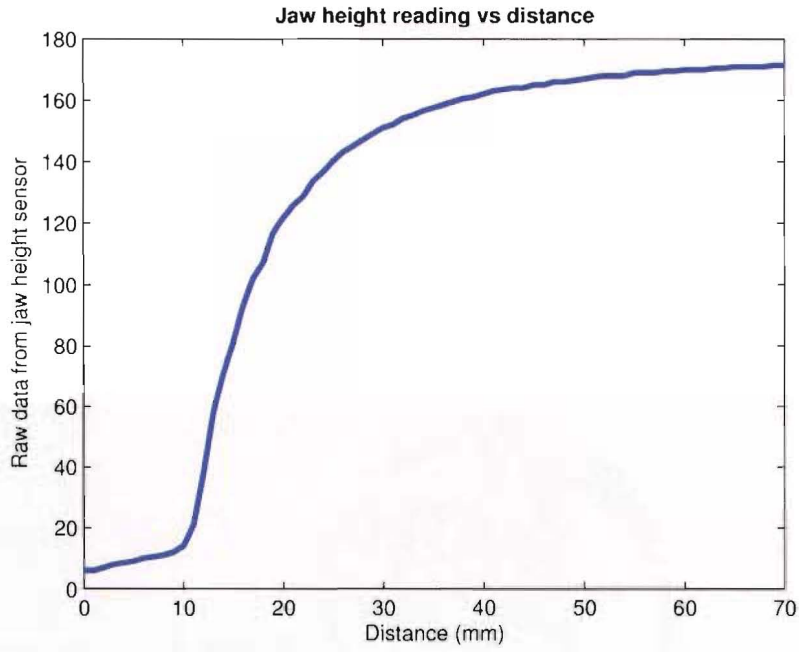

Figure 8.9 The jaw height sensor output versus distance.

click on the 'Start' button to begin using the device. When in use, the function *Jaw2Freq* takes iInput from the serial port, m_Max, m_Min and Habitual Pitch (from the initial setup) to convert iInput to *F0* based on the fact that *F0* varies by ± 3.4 semitones from the habitual pitch during normal conversation. If iInput is of the same value for 150 samples or more (e.g. the user leaves his/her mouth in the same position for 0.5 second or more), *F0* is set to 0 (e.g. the artificial glottal source stops vibrating). As the user opens his/her mouth to a certain threshold value (the default on threshold value is 10 ADC units which corresponds to $\approx$ 5mm of jaw height movement), the glottal sound source begins to vibrate again. The sensitivity of the device can be reduced by increasing the threshold value. Refer to the flow diagram of the *Jaw2Freq* function in Figure 8.11.

The next stage is the glottal pulse generator using the twin-bar model. A function *TBModel* is called with *F0* as input and an optional gender input. The output from this function is the gTB, the glottal pulse generated with the twin-bar model. Refer to section 5.3.3 for the equations involved in the generation of the twin-bar glottal pulse.

From here, the glottal pulse goes through the trachea filter (*TracheaFilt*) to compensate for the shift in location of the glottal source. The reason for this is that the glottal source for a normal person is inside the larynx, but with this artificial device, the source can either be in the mouth cavity (default) or in the pharynx. The function *TracheaFilt* is called with gTB as input and, source location as an optional input (since the default source location is inside the mouth cavity). The process for

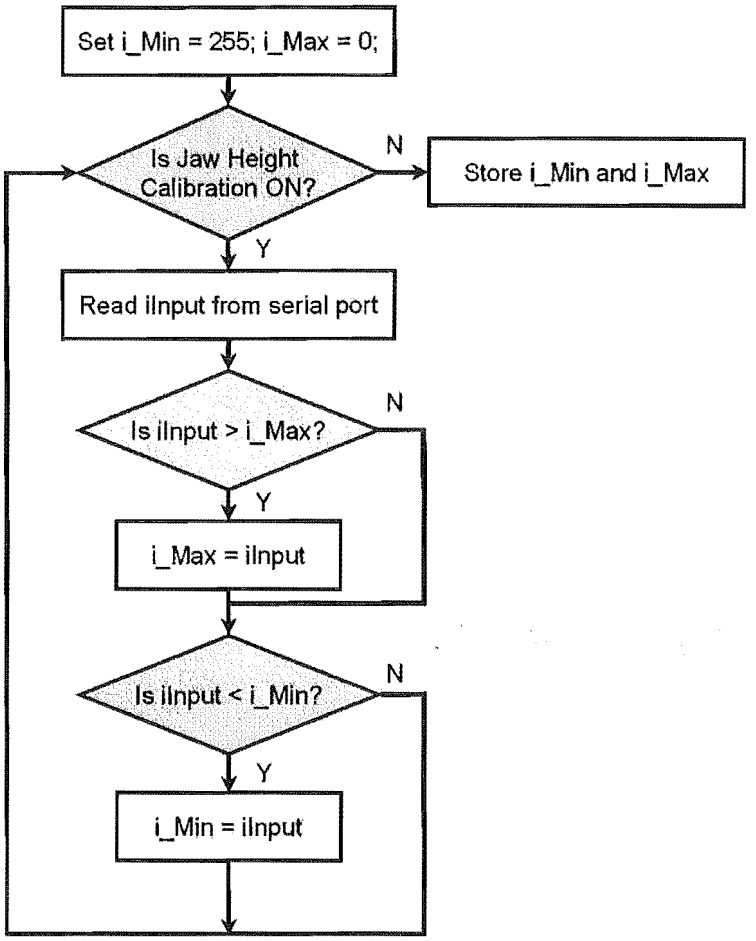

Figure 8.10 The signal flow chart for the jaw height calibration.

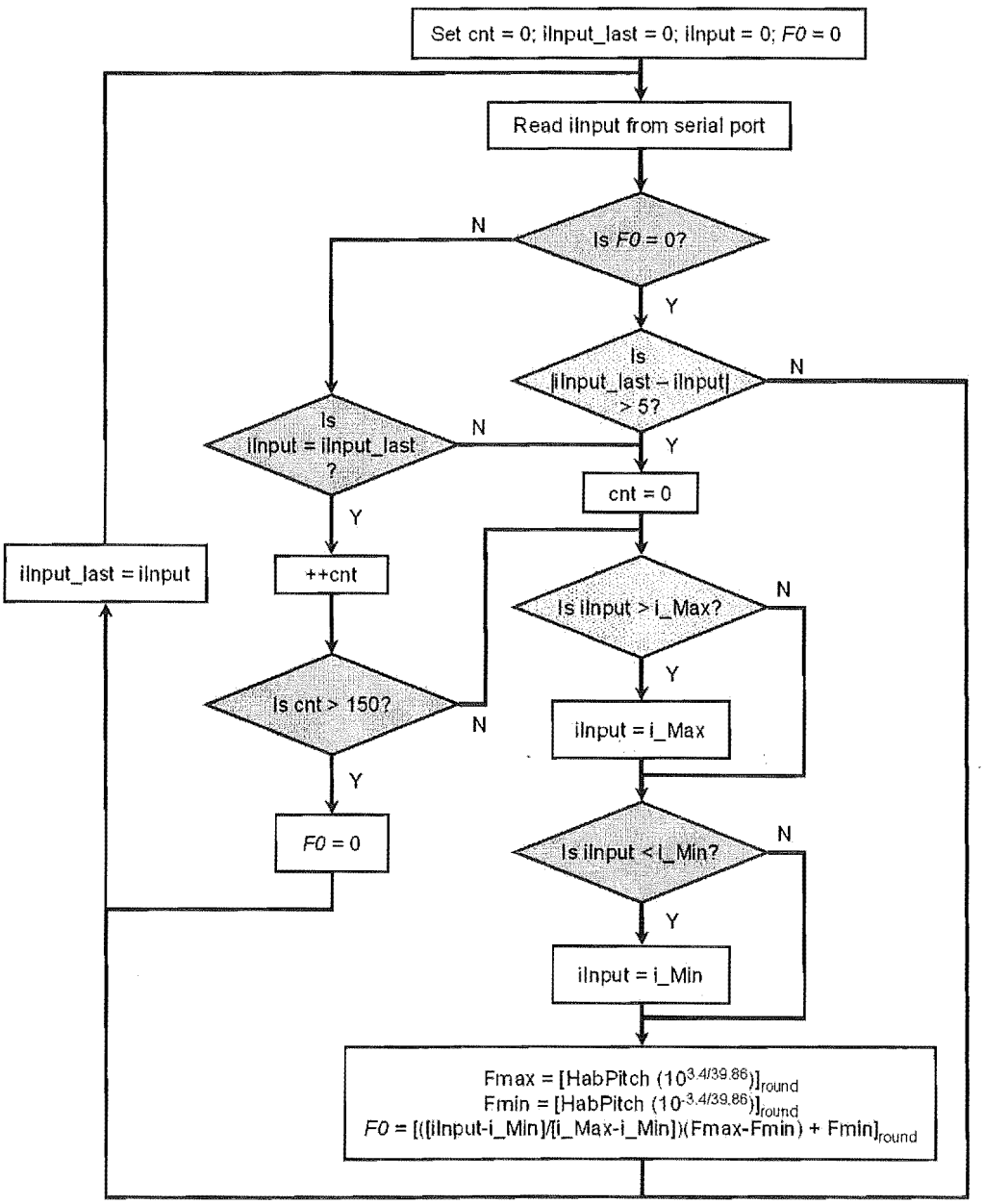

Figure 8.11 The signal flow chart for converting jaw height to *F0* (*Jaw2Freq*) and automatic on/off switching of the sound source generator.

filtering the gTB is the same as in section 7.2 on voice synthesis. The section of vocal tract between the vocal folds and the new location of the artificial sound source is modelled using the half-sample delay Kelly-Lochbaum structure as shown in section 5.4.3. The only difference is that the vocal tract length is now 47mm (6 segments of the vocal tract area function shown in Figure 7.11) if the source is located in the pharynx, or 94mm (12 segments) if the source is located in the mouth cavity, instead of the normal 21 or 22 segments used in voice synthesis. The area of each segment is obtained from the average vocal tract area function of the 6 vowels used for the waveform shape experiment in section 7.1. The output of the *TracheaFilt* function is FiltgTB, the actual sound source that is to be sent to the BT headset.

The glottal pulse FiltgTB is sent to the BT (hearing aid) speaker via the BT USB dongle. For this prototype, the default sound source is the mouth cavity. The speaker is therefore covered with a short intra-oral tube to prevent saliva from flowing into the speaker. In practice, the speaker will have to be modified to make it water-proof and the transfer function of this tube will have to be taken into account. With the sound source inside, the artificial voice, UnfiltV, is generated as the user "mouths the words". The BT microphone detects this voice and sends it back to the laptop via the BT USB dongle.

Inside the laptop, the speech enhancement process is carried out to remove unwanted noise and also to amplify the UnfiltV. The function *SpEnhance* is called with UnfiltV as input and Voice as output. Voice is then send to the laptop sound card and speaker where it is perceived as speech.

8.3.1 Prototype testing

A preliminary test was carried out to test the voice quality of *MyVoice2* (without the speech enhancement stage) compared with an intraoral electrolarynx (shown in Figure 3.6). With an audience of 3 people, the author uttered short sentences containing voiced sounds (e.g. "Hello, how are you?" and "My name is Marilyn.") with *MyVoice2* and then with the electrolarynx. All parties agreed that *MyVoice2* produced an improved voice quality compared with the electrolarynx. However, the volume of the voice produced with *MyVoice2* was a bit soft due to the small size of the hearing aid speaker. The hearing aid speaker was later replaced with a bigger speaker placed inside a mouth-guard, as shown in the sample clips of voices generated by *MyVoice2* and an electrolarynx on the CD at the back of this thesis.

The jaw sensor, due to the way it has been mounted on the head, is quite sensitive to jaw movement. Whenever the head or jaw moves, the jaw sensor shifts as well. As a result of that, it is sometimes difficult for the glottal source to turn off. The headset had to be mounted slightly above the ear to reduce this artifact. A new wireless mouth-guard prototype will help prevent this problem (see section 9.2.2 on future work for more details).

Chapter 9

Conclusion and suggestions for future research

The final chapter of this thesis presents the conclusion for the research involved in the design and build of an artificial speech device for speech impaired individuals (section 9.1) and to discuss the possible options for future research on this topic (section 9.2).

9.1 Conclusion

The ultimate aim of this research is to create an artificial speech device that will allow patients in intensive care unit (ICU) and laryngectomees who are unable to speak due to airway obstruction to produce natural sounding speech by "mouthing the words". Current artificial speech devices are not suitable for ICU patient and/or laryngectomees for a number of reasons. Some are difficult to operate while others, the generated voice does not sound natural. Some communication techniques are not efficient, they consume too much time and effort to convey a simple message. And then there are others that take a lot of time to learn.

The naturalness of speech is affected by a number of factors. In the design of artificial speech device, the two key elements are: pitch variation and glottal sound source. Pitch variation of a normal person's voice originates from the vibrating vocal folds. In ICU patients and laryngectomees, the vocal folds are either non-functional or do not exist at all. This research looked at the alternative methods for controlling pitch variation of an artificial speech device by using other part(s) of the body. An obvious choice for this is to use jaw height (since a person's jaw automatically moves as the person speaks). A study was carried out to determine the relationship between jaw height and pitch. Results from the preliminary study suggest that jaw height is negatively proportional to *F0*. A non-contact jaw height detector (pitch controller) using reflective object sensor was later designed

as a result of this study.

The electroglottograph (EGG) signal was employed as the choice for glottal sound source for this thesis as it is easy to measure, readily available, independent of vowel effect and has a similar shape to glottal airflow (sometimes used as the glottal sound source for artificial speech simulation). An EGG analysis method was developed to find the average EGG waveform for a given utterance. The EGG analysis software is also used to extract the key features of the average EGG waveform (glottal pulse), e.g. *F0*, *OP*, *CP* and *sCP*, that provide markers for determining the shape of a glottal pulse.

Another study was carried out to find the relationship between glottal pulse shape and *F0*. The parameters *OP*, *CP* and *sCP* were all found to decrease with *F0*. A new glottal pulse model, the twin-bar model was created through the results obtained from this study. The unique feature of the twin-bar model is that the shape of the glottal pulse changes with respect to a single parameter - *F0*. Other parameters involved in the glottal waveform synthesis are either predetermined or can be calculated with *F0* as input.

A third and final study was carried out to test the quality of synthesised voice produced using the twin-bar model, the LF glottal model and the Rosenberg's glottal model. Results from the perceptual study showed that synthesised voice generated with the twin-bar model is significantly better than the other two glottal models.

With the information gathered from all three studies, a prototype artificial speech device, known as *MyVoice*, was developed. Preliminary test of the device on a normal subject showed positive results.

This research achieved its goals. The prototype (*MyVoice2*), although cumbersome at this stage, proved that it is possible to vary pitch automatically as the subject "mouths the words", the voice sounded more natural using the twin-bar model as glottal sound source and the device is user friendly - its operation is intuitive and therefore requires minimal learning experience.

9.2 Suggestions for future research

There are a number of aspects of the research presented in this thesis that require further investigation.

9.2.1 Experiments and glottal model

The studies conducted in this research were limited to either a single gender or a single nationality and in a single language. It may not be accurate to apply the results of these experiments to the general population, especially when some languages are tonal for which the relationship between jaw height and *F0* may not exist. It is suggest that further studies should include both male and

female subjects from different age groups, nationalities and languages where the data from each group is analysed separately.

The study of the waveform shape and the jaw height variation with *F0* should also include different modes of phonation, such as vocal fry and falsetto, instead of having the twin-bar glottal sound source that only operates in a single voice mode; the modal register. Perhaps the top bar should have two separate bars at a fixed distance apart for generating the two peaks on a single glottal pulse in vocal fry mode. For falsetto mode, it may be possible to change the slope of the returning phase of the bottom bar so that the glottal pulse becomes more symmetrical (a characteristic of the falsetto mode).

The other aspects of natural speech, such as voice intensity, were not covered in this thesis. If there is a relationship between *F0* and voice intensity, it may be incorporate in the twin-bar model to further improve the naturalness of voice generated with this glottal model.

The current design for *MyVoice2* is only designed to generate voiced sounds. The artificial voice produced at the lips may however contain some unvoiced sounds. For example, a subject "mouthing" the word 'pot' may produce a light 'pop' sound that can be picked up by the microphone. It would be interesting to find out whether it is possible for unvoiced sounds (e.g. /p/, /t/, /s/ and /f/) to be artificially generated in the speech enhancement stage and incorporated to improve the intelligibility of the speech produced.

The number of subjects participated in the jaw height and vocal folds movement experiment was limited. A few more subjects are required to establish a more accurate measure of the the relationship of jaw height with *F0*. Similarly, more subjects are needed for glottal waveform shape experiment to measure the EGG waveform shape so that a more accurate glottal pulse model can be obtained.

An assumption made in this thesis is that the EGG can represent the glottal sound source because its characteristics resemble the glottal airflow signal. It will be interesting to find out the quality of the synthesised voice generated with the twin-bar model using glottal airflow and inverse filtered acoustic signal, and to compare them with those generated using the EGG signal.

To evaluate the effectiveness of the prototype artificial speech device (*MyVoice2*), the first step is to test the quality of the artificial voice generated with *MyVoice2* compared with the voice generated with an electrolarynx on a normal subject or a laryngectomee. Sounds to be tested should include isolated vowels, short sentences and short conversations. Volunteers will then be required to listen to these test sounds to decide which of the two devices produce a better or more natural sounding voice. If the results of this evaluation looks promising, then *MyVoice3*, a totally wireless mouth-guard version (see next section) should be built before the evaluation of the effectiveness of the artificial voice device can be extended to proper clinical trial.

9.2.2 Hardware development

The default minimum frequency for *MyVoice2* was set at above 100Hz since small speakers are known to have poor low frequency response. A new design should take into account the transfer functions of the speakers and microphone to ensure that the desired signal is attained.

A fully wireless headset using radio-frequency (RF) transmitter-receiver set (e.g. TWS-BS3 & RWS-371-6 from Wenshing Electronics Co., Ltd.) for the jaw height detector or Bluetooth module (Promi-ESD-02) for jaw height and glottal pulse transmission and the use of rechargeable batteries for the power supply will make the design more compact. The headset currently is wireless for the hearing aid speaker and microphone but not for the jaw height detector.

One downside of *MyVoice2* is that the jaw height sensor needs to be positioned properly otherwise sufficient pitch variation may not be acquired. One solution to this is to place the non-contact jaw height sensor on top of the shirt collar, similar to a microphone, so that it can be discreet. Moving the head may then change the pitch but if the subject does not open his/her mouth the voice is not radiated and may sound like a very weak hum. The subject will just have to make sure that the head is in its neutral position when he/she wishes to speak to obtain the desired pitch variation. This option is probably better for laryngectomees than for ICU patients as patients in hospital usually wear hospital gowns with no collar. The headset option is still the better option for ICU patients.

An even better option is to change the design of the jaw height sensor to one that is molded inside a mouth-guard (e.g. use a surface mount infrared (IR) emitter and phototransistor pair instead of the larger IR reflective object sensor). This is because the ventilator lines around the throat region of the ICU patients may interfere with the operation of the current (external) jaw height detector.

In the case of the in pharynx operation: a custom designed tracheal tube/nasogastric tube that will allow the insertion of a speaker and a microphone while still allowing the full function of the tube is desirable.

In the mouth cavity operation: a more comfortable mouth piece will be an advantage, possibly a custom-made small thin mouth-guard with the hearing aid speaker, jaw height detector, Bluetooth transceiver unit and rechargeable battery molded inside. The mouth piece needs to be made water-proof so that the user does not have to take the mouth piece off to drink. When a patient wants to use the device, the caregiver can place the water-proof device into a warm cup of water and when the mouth-guard is soft, place it inside the subject's mouth and mold it into shape. This is far easier and quicker to make than a custom-made retainer.

The prototype is quite bulky at the moment. A suggestion for the next version is to use a digital signal processor (DSP) for the signal processing thereby removing the need for the laptop as the main processing unit and making the device more compact and portable. To improve speech quality,

a better speech enhancement algorithm that includes ambient noise cancellation (e.g. reducing noise produced by machines in the ICU) and unvoiced sounds (based on the signal picked up by the microphone) should be included as well.

References

[ABT00] F. Alipour, D. A. Berry, and I. R. Titze. "A finite-element model of vocal fold vibration," *J. Acous. Soc. Am.*, vol. 108, pp. 3003–3012, 2000.

[Arn61] G. E. Arnold. "Physiology and pathology of the cricothyroid muscle," *Laryngoscope*, vol. 71, pp. 687–753, July 1961.

[Ava02] F. Avanzini. *Computational Issues in Physically-based Sound Models.* Ph.D. thesis, Dottorato di ricerca in ingegneria informatica ed elettronica industriali, Università degli Studi di Padova Dipartimento di Elettronica ed Informatica, 2002.

[Bau00] N. Bauman. "http://www.saywhatclub.com/newsletter/mar00/neil.htm," March 2000.

[BB98] J. E. Bernthal and N. W. Bankson. *Articulation and phonological disorders.* Allyn and Bacon, 4th edition, 1998.

[BCNG98] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert. "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *J. Acous. Soc. Am.*, vol. 103, no. 5, pp. 2649–2658, May 1998.

[Bea68] W. P. Beatrous. "Tracheostomy (tracheotomy). Its expanded indications and its present status. Based on an analysis of 1,000 consecutive operations and A review of the recent literature," *Laryngoscope*, vol. 78, no. 1, pp. 3–55, 1968.

[BEH89] I. Bergbom-Engberg and H. Haljamae. "Assessment of patients' experience of discomforts during respirator therapy," *Crit. Care Med.*, vol. 17, no. 10, pp. 1068–1072, October 1989.

[BF04] C. Byrne and P. Foulkes. "The 'mobile phone effect' on vowel formants," *Speech, Language and the Law, University of Birmingham Press*, vol. 11, no. 1, pp. 83–102, 2004.

[Blo78] E. D. Blom. "The artificial larynx: Past and present," in *The Artificial Larynx Handbook*, S. J. Salmon and L. P. Goldstein, Eds. 1978, p. 57, Grune and Stratton, New York.

[BO00] R. J. Baken and Robert F. Orlikoff. *Clinical Measurement of Speech and Voice*. Thomson Learning TM, Singular Publishing Group, 2nd edition, 2000.

[Bos94] Z. T. Bosone. *Laryngectomee Rehabilitation*, chapter Tracheoesophageal fistulization/puncture for voice restoration: Pre-surgery considerations and troubleshooting procedures, p. 359. PRO-ED, Austin, TX, 3rd edition, 1994.

[Bro98] M. Brookes. "http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html," 1998.

[BYNG98] M. Blomgren, C. Yang, M. L. Ng., and H. R. Gilbert. "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *J. Acous. Soc. Am.*, vol. 103, no. 5, pp. 2649–2658, 1998.

[CBC+02] D. Cook, R. Brower, J. Cooper, L. Brochard, and J. Vincent. "Multicenter clinical research in adult critical care," *Critical Care Medicine*, vol. 30, no. 7, pp. 1636–1643, July 2002.

[CC93] J. K. Casper and R. H. Colton. *Clinical manual for laryngectomy and head and neck cancer rehabilitation*. Singular, San Diego, 1993.

[CC96] R. H. Colton and J. K. Casper. *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment*. Williams & Wilkins, Baltimore, 2nd edition, c1996.

[Chi84] D. G. Childers. "A critical review of electroglottography," *CRC Critical Review in Biomedical Engineering*, vol. 12, no. 2, pp. 131–161, 1984.

[CHMA86] D. G. Childers, D. M. Hicks, G. P. Moore, and Y. A. Alsaka. "A model for vocal fold vibratory motion, contract area, and the electroglottogram," *J. Acous. Soc. Am.*, vol. 80, no. 5, pp. 1309–1320, 1986.

[CL91] D. G. Childers and C. K. Lee. "Vocal quality factors: Analysis, synthesis and perception," *J. Acous. Soc. Am.*, pp. 2394–2410, 1991.

[CP86] A. Cutler and M. Pearson. *Intonation in Discourse*, chapter On the analysis of prosodic turn-taking cues, pp. 139–155. Croom Helm, London, 1986.

[CW90] D. G. Childers and K. Wu. "Quality of speech produced by analysis-synthesis," *Speech Comm.*, vol. 9, pp. 97–117, 1990.

[dB58] J. W. Van den Berg. "Myoelastic-aerodynamic theory of voice production," *J. Speech Hear. Res.*, vol. 1, pp. 227–244, 1958.

[Del83] J. R. Deller. "On the time domain properties of the two-pole model of the glottal waveform and implications for LPC," *Speech Communication: An Interdisciplinary Journal*, vol. 2, pp. 57–63, 1983.

[DH02] J. Dang and K. Honda. "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," *Journal of Phonetics*, 2002.

[Die91] W. M. Diedrich. "Anatomy and physiology of esophageal speech," in *Alaryngeal Speech Rehabilitation for Clinicians by Clinicians*, S. J. Salmon and K. H. Mount, Eds. 1991, p. 2, PRO-ED, Austin, TX.

[DM04] B. J. Daley and M. Munyikwa. "http://www.emedicine.com/med/topic2975.htm," 2004.

[Doy94] P. C. Doyle. *Foundations of Voice and Speech Rehabilitation Following Laryngeal Cancer*. San Diego: Singular, 1994.

[DY66] W. M. Diedrich and K. A. Youngstrom. *Alaryngeal Speech*. Springfield, IL: Charles C. Thomas, 1966.

[Ede83] Y. Edels. "Pseudo-voice: Its theory and practice," in *Laryngectomy: Diagnosis to Rehabilitation*, Y. Edels, Ed. 1983, p. 112, Aspen, Rockville, MD.

[EFP98] D. Erickson, O. Fujimura, and B. Pardo. "Articulatory correlates of prosodic control: Emphasis and emotion," *Lang. and Speech*, vol. 41, pp. 399–417, 1998.

[EH96] D. Erickson and K. Honda, Eds. *Jaw displacement and F0 in contrastive emphasis*. ASA 131st Meeting, Indianapolis, May 1996.

[EIEF04] D. Erickson, R. Iwata, M. Endo, and A. Fujino. "Effect of tone height on jaw and tongue articulation in Mandarin Chinese," in *International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages, Beijing, China*, W. Hess, Ed., 28-31 March 2004.

[Ele] Kay Elemetrics. "Model 6103," *http://www.kayelemetrics.com*.

[EM03] I. Eliachar and C. Milstein. "Hands-free speech in long-term tube-free tracheostomy," in *32nd Annual Symposium: Care of the Professional Voice, The Voice Foundation*, C. N. Ford, R. T. Sataloff, J. A. Koufman, P. Woo, C. Rosen, and S. M. Zeitels, Eds., 4-8 June 2003.

[EMHD00] D. Erickson, K. Maekawa, M. Hashi, and J. Dang. "Some articulatory and acoustic changes associated with emphasis in spoken English," *Proceedings of the International Conference of Spoken Language Processing*, vol. 3, pp. 247–250, 2000.

[Ewa79] W. G. Ewan. "Can intrinsic vowel f0 be explained by source or tract coupling?," *J. Acous. Soc. Am.*, vol. 66, pp. 358–362, 1979.

[FA57] K. Faaborg-Anderson. "Electromyographic investigation of intrinsic laryngeal muscles in humans," *Acta Physiol. Scand. Suppl.*, , no. 41, pp. 140, 1957.

[Fai60] G. Fairbanks. *Voice and Articulation Drillbook.* Harper and Row, New York, 1960.

[Fan60] G. Fant. *Acoustic Theory of Speech Production.* Mouton, Hague, 1960.

[Fan73] G. Fant. *Speech Sounds and Features.* MIT Press, Cambridge, Massachusetts, 1973.

[FL68] J. L. Flanagan and L. Landgraf. "Self-oscillating source for vocal tract synthesizers," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 57–64, 1968.

[Fla57] J. L. Flanagan. "Note on the design of terminal-analog speech synthesizers," *J. Acous. Soc. Am.*, vol. 29, pp. 306–310, 1957.

[Fla72] J. L. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer Verlag, New York, 2nd edition, 1972.

[FLL85] G. Fant, J. Liljencrants, and Q. G. Lin. "A four-parameter model of glottal volume-velocity," *STL_QPSR, Royal Institute of Technology, Stockholm, Sweden*, vol. 4, pp. 1–13, 1985.

[FR91] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments.* Springer Verlag, New York, 1991.

[Fry79] D.B. Fry. *Physics of Speech.* Cambridge University Press, 4th edition, 1979.

[Geu01] A. Geumann. "Vocal intensity: Acoustic and articulatory correlates," in *Proc. of the 4th International Speech Motor Conference, Nijmegen*, H. F. M. Peters, Ed., 13-16 June 2001.

[GH61] W. N. Gardner and H. E. Harris. "Aids and devices for laryngectomees," *Arch. Otolaryngol*, vol. 73, pp. 145–152, February 1961.

[Gil94] S. I. Gilmore. *Laryngectomee Rehabilitation*, chapter The physical, social, occupational, and psychological concomitant of laryngectomy, p. 400. PRO-ED, Austin, TX, 3rd edition, 1994.

[Gor96] E. Gordon. *Phonemic Transcription Workbook.* Department of Linguistics, University of Canterbury, New Zealand, 1996.

[Gra97] M. S. Graham. *The Clinician's Guide to Alaryngeal Speech Therapy.* Butterworth-Heinemann, 1997.

[GRJ+82] G. A. Gates, W. Ryan, J. C. Cooper Jr., G. F. Lawlis, E. Cantu, T. Hayashi, E. Lauder, R. W. Welch, and E. Hearne. "Current status of laryngectomee rehabilitation: I. Results of therapy," *Am. J. Otolaryngol*, vol. 3, no. 1, pp. 1–7, January-February 1982.

[Har75] J. Harden. "Comparison of glottal area changes as measured from ultra-high-speed photographs and photoelectric glottographs," *J. Speech and Hear. Res.*, vol. 18, pp. 728–738, 1975.

[HHKM99] K. M. Houston, R. E. Hillman, J. B. Kobler, and G. S. Meltzner. "Development of sound source components for a new electrolarynx speech prosthesis," in *24th International Conference on Acoustics, Speech and Signal Processing (ICASSP), Sponsored by: IEEE*, 15-19 March 1999.

[Hir74] M. Hirano. "Morphological structure of the vocal cord as a vibrator and its variations," *Folia Phoniatr.*, vol. 26, pp. 89–94, 1974.

[Hir88] H. Hirose. "High-speed digital imaging of vocal fold vibration," *Acta Otolaryngol. Suppl.*, vol. 458, pp. 151–153, 1988.

[HJ73] H. Hollien and B. Jackson. "Normative data on the speaking fundamental frequency characteristics of young adult males," *Journal of Phonetics*, vol. 1, pp. 117–120, 1973.

[Hol68] H. Hollien. "Perceptual study of vocal fry," *J. Acous. Soc. Am.*, vol. 43, no. 3, pp. 506–509, 1968.

[Hol73] J. N. Holmes. "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 298–305, 1973.

[Hol74] H. Hollien. "On vocal registers," *J. Phon.*, vol. 2, pp. 25–43, 1974.

[HOV69] M. Hirano, J. Ohala, and W. Vennard. "The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation," *J. Speech Hear. Res.*, , no. 12, pp. 616–628, 1969.

[HRC03] N. Henrich, B. Roubeau, and M. Castellengo. "On the use of electroglottography for characterisation of the laryngeal mechanisms," *Proceedings of the Stockholm Music Acoustics Conference (SMAC), Sweden*, vol. SMAC-1, 6-9 August 2003.

[HSLDK95] D. J. Higginbotham, R. Sonnenmeier, S. Lawrence-Dederich, and K. Kim. "Assistive technologies for disorders of expressive communication and cognition," in *Assistive Communication for Persons with Disabilities*, W. Mann and J. Lane, Eds. American Occupational Therapy Association, Rockville, MD., 1995.

[IES05] IES. "http://www.cavs.msstate.edu/hse/ies/publications/courses/ase_6713/lecture_02.pdf," *Department of Electrical and Computer Engineering, Mississippi State University*, 2005.

[IF72] K. Ishizaka and J. L. Flanagan. "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 512, pp. 1233–1268, 1972.

[III94] W. T. S. Fitch III. *Vocal Tract Length Perception and the Evolution of Language*. Ph.D. thesis, The Department of Cognitive and Linguistic Sciences, Brown University, 1994.

[JABM87] H. R. Javkin, N. Antonanzas-Barosso, and I. Maddieson. "Digital inverse filtering for linguistic research," *J. Speech and Hear. Res.*, vol. 30, pp. 122–129, 1987.

[JHP00] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis. *Discrete-time Processing of Speech Signals*. IEEE Press, 2000.

[JJ98] A. F. Johnson and B. H. Jacobson. *Medical Speech-language Pathology: A Practitioner's Guide*. Thieme Medical Publishers, New York, 1998.

[JL86] C. John-Lewis. *Intonation in Discourse*, chapter Prosodic differentiation of discourse modes, pp. 199–219. Croom Helm, London, 1986.

[KAR99] M. Kob, N. Alhäuser, and U. Reiter. "Time-domain model of the singing voice," in *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), NTNU, Trondheim*, 9-11 December 1999, pp. W99–1–4.

[Kei74] R. L. Keith. *A Handbook for the Laryngectomee*. Interstate printers and publishers, Inc., 1974.

[KFP68] P. S. King, E. W. Fowlks, and G. A. Pierson. "Rehabilitation and adaptation of laryngectomy patients," *Am. J. Phys. Med.*, vol. 47, pp. 192, 1968.

[KL62] J. L. Kelly and C. C. Lochbaum. "Speech synthesis," in *Proc. of the 4th International Congress on Acoustics, Paper G42*, September 1962, pp. 1–4.

[LB88] P. Lieberman and S. E. Blumstein. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge studies in speech science and communication, Cambridge University Press, Cambridge, U.K., 1988.

[Ler91] J. W. Lerman. "The artificial larynx," in *Alaryngeal Speech Rehabilitation for Clinicians by Clinicians*, S. J. Salmon and K. H. Mount, Eds. 1991, p. 29, PRO-ED, Austin, TX.

[Lie67] P. Lieberman. *Intonation, Perception and Language*. MIT Press, Cambridge MA, 1967.

[Lin85] R. Linggard. *Electronic Synthesis of Speech*. Cambridge University Press, Cambridge, U.K., 1985.

[LJNH00] E. Lin, J. Jiang, S. D. Noon, and D. G. Hanson. "Effects of head extension and tongue protrusion on voice perturbation measures," *J. Voice*, vol. 14, pp. 8–16, 2000.

[LMJ88] P. Ladefoged, I. Maddieson, and M. Jackson. "Investigating phonation types in different languages," in *Vocal fold Physiology: Voice Production, Mechanisms and Functions*, O. Fujmura, Ed. 1988, pp. 297–317, Raven Press, New York.

[Luc04] J. C. Lucero. "Dynamics of the vocal fold oscillation," in *XXVII National Congress on Applied and Computational Mathematics - CNMAC, Porto Alegre*, 13-17 September 2004.

[Mac82] M. Maclagan. "An acoustic study of New Zealand English vowels," *The New Zealand Speech Therapists Journal*, vol. 37, pp. 20–26, 1982.

[Mac00] M. Maclagan. "Acoustic cues for vowels and consonants (lecture handout)," *Christchurch Teachers College, Speech Therapy Department*, 2000.

[Mae76] S. Maeda. *A Characterization of American English Intonation*. Ph.D. thesis, MIT, 1976.

[Mar94a] N. Maragos. *Laryngectomee Rehabilitation*, chapter Anatomy and physiology of the laryngecotmee, p. 72. PRO-ED, Austin, TX., 1994.

[Mar94b] D. E. Martin. *Darley Laryngectomee Rehabilitation*, chapter Evaluating esophageal speech development and proficiency, p. 334. PRO-ED, Austin, TX, 3rd edition, 1994.

[Mar94c] D. E. Martin. *Laryngectomee Rehabilitation*, chapter Pre- and postoperative anatomical and physiological observations in laryngectomy, p. 79. PRO-ED, Austin, TX, 3rd edition, 1994.

[Mar96] K. Marasek. "Glottal correlates of the word stress and the tense or lax opposition in the German vowels," *Proc. of the ICSLP-96*, pp. 1573–1577, 1996.

[Mas93] M. F. Mason. *Speech Pathology for Tracheostomized and Ventilator Dependent Patients*. Voicing! Inc., Newport Beach, California, 1993.

[MG76] J. D. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer Verlag, New York, 1976.

[MHG96] D. Maurer, M. Hess, and M. Gross. "High-speed imaging of vocal fold vibrations and larynx movements within vocalizations of different vowels," *Ann. Otol. Rhhinol. Laryngol.*, vol. 105, pp. 975–981, 1996.

[ML90] D. Molyneaux and V. W. Lane. *Successful Interactive Skills for Speech-language Pathologists and Audiologists*. Rockville, MD: Aspen, 1990.

[Möb03] B. Möbius. "Gestalt psychology meets phonetics - an early experimental study of intrinsic F0 and intensity," *15th ICPhS Barcelona*, pp. 2677–2680, 2003.

[Moo68] G. P. Moore. "Otolaryngology and speech pathology," *Laryngoscope*, vol. 78, pp. 1500–1507, 1968.

[Moo75] G. P. Moore. "Voice problems following limited surgical excision," *The laryngoscope*, vol. 85, no. 4, pp. 619–625, April 1975.

[MvL58] G. P. Moore and H. S. von Leden. "Dynamic variation of the vibratory pattern in the normal larynx," *Folia Phoniatrica*, vol. 10, pp. 205–238, 1958.

[OS75] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, New Jersey, 1975.

[Pet78] N. Petersen. "Intrinsic fundamental frequency of Danish vowels," *J. Phonetics*, vol. 6, pp. 177–189, 1978.

[PGU98] A. W. Pearl, P. J. Gannon, and M. L. Urken. "Anatomy and vascular perfusion territories of the superior thyroid artery in Macaca mulatta," *Laryngoscope*, July 1998.

[Pic98] J. M. Pickett. *The Acoustic of Speech Communication: Fundamentals, Speech Perception Theory, and Technology*. Allyn & Bacon, MA., 1998.

[Pul05] H. Pulakka. *Analysis of Human Voice Production using Inverse Filtering, High-Speed Imaging, and Electroglottography*. Ph.D. thesis, Department of Computer Science and Engineering, Helsinki University Of Technology, 2005.

[Put61] E. J. Putney. "Rehabilitation of the post-laryngectomized patient," *Arch. Otolaryngol*, vol. 73, pp. 145, 1961.

[RE96] C. Van Riper and R. L. Erickson. *Speech Correction: An Introduction to Speech Pathology and Audiology*. Allyn and Bacon, 9 edition, 1996.

[RFBS84] J. Robbins, H. B. Fisher, E. D. Blom, and M. I. Singer. "A comparative acoustic study or normal, esophageal, and tracheoesophageal speech production," *J. Speech Hear. Disord.*, , no. 49, pp. 202, 1984.

[RHI05] D. Rice, A. Harnett, and P. Inkpen. "Structure of the human respiratory system. http://www.cdli.ca/ dpower/biology.htm," 2005.

[Rob94] N. K. Roberts. *Laryngectomee Rehabilitation*, chapter Nursing intervention for the laryngectomee: Management of change in self-care practices following hospitalization, p. 121. PRO-ED, Austin, TX, 3rd edition, 1994.

[Ros71] A. E. Rosenberg. "The effect of glottal pulse shape on the quality of natural vowels," *J. Acous. Soc. Am.*, vol. 49, no. 2, pp. 583–590, 1971.

[Rot73] M. Rothenberg. "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acous. Soc. Am.*, vol. 53, pp. 1632–1645, 1973.

[Sal83] S. J. Salmon. "Artificial larynx speech: A viable means of alaryngeal communication," in *Laryngectomy: Diagnosis to rehabilitation*, Y. Edels, Ed. 1983, p. 143, Aspen, Rockville, MD.

[Sal86a] S. J. Salmon. *Laryngectomee Rehabilitation*, chapter Laryngectomee visitations, p. 149. College-Hill, San Diego, 2nd edition, 1986.

[Sal86b] S. J. Salmon. *Pre- and postoperative conferences with laryngectomees and their spouses. In: RL Keith, FL Darley (Eds), Laryngectomee rehabilitation (2nd Ed).* San Diego: College-Hill, 1986.

[SB80] M. I. Singer and E. D. Blom. "An endoscopic technique for restoration of voice after laryngectomy," *Ann. Otol. Rhinol. Laryngol.*, vol. 89, no. 6, pp. 529–533, November-December 1980.

[SBH81] M. I. Singer, E. D. Blom, and R. C. Hamaker. "Further experience with voice restoration after total laryngectomy," *Ann. Otol. Rhinol. Laryngol.*, vol. 90, no. 5, pp. 498–502, September-October 1981.

[Sen] Sensimetrics. "Speechstation2," *http://www.sens.com/speechstation_overview.htm.*

[Ser02] Tech Connections. Assistive Technology Quick Reference Series. "http://www.techconnections.org/resources/guides/commdevices.pdf," 2002.

[SH72] T. Shipp and R. M. Haller. "Vertical larynx height during vocal frequency change," in *Paper presented at 83rd ASA meeting, Buffalo*, 1972.

[Ske79] M. Skelly. *Amerind Gestural Code Based on Universal American Indian Hand Talk.* Elsevier North Holland, New York., 1979.

[sl00] Computerized speech lab. "Model 4300b. version 4.3," *http://www.kayelemetrics.com*, 2000.

[Soc95] American Cancer Society. "First steps: Helping words for the laryngectomee," *Atlanta: American Cancer Society*, p. 14, 1995.

[Son86] M. M. Sondhi. "Resonances of a bent vocal tract," *J. Acous. Soc. Am.*, vol. 79, pp. 1113–1116, April 1986.

[SS78] P. H. Skinner and R. L. Shelton. *Speech, Language and Hearing: Normal Processes and Disorders.* Addison-Wesley Publishing Company, 1978.

[SS94] J. Schroeter and M. M. Sondhi. "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE transactions on speech and audio processing*, pp. 133–150, January 1994.

[ST95] B. H. Story and I. R. Titze. "Voice simulation with a body-cover model of the vocal folds," *J. Acous. Soc. Am.*, vol. 97, pp. 1249–1260, 1995.

[ST96] B. H. Story and I. R. Titze. "Vocal tract area functions from magnetic resonance imaging," *J. Acous. Soc. Am.*, vol. 100, no. 1, pp. 537–553, July 1996.

[Sto02] B. H. Story. "An overview of the physiology, physics and modeling of the sound source for vowels," *Acoustical Science and Technology*, vol. 23, no. 4, pp. 195–206, 2002.

[Sun73] J. Sunners. "The use of the electrolarynx in patients with temporary tracheostomies," *J. Speech Hear. Dis.*, , no. 38, pp. 335–338, 1973.

[TE93] H. Traunmüller and A. Eriksson. "The frequency range of the voice fundamental in the speech of male and female adults," *Institutionen for lingvistik, Stockolms universitrt, S-106-91 Stockholm, Sweden*, (manuscript) 1993.

[Tip00] D. C. Tippett. *Tracheostomy and Ventilator Dependency: Management of Breathing, Speaking, and Swallowing*. Thieme Medical Publishers, Inc., 2000.

[Tit73] I. R. Titze. "The human vocal cords: A mathematical model, Part I," *Phonetica*, vol. 28, pp. 129–170, 1973.

[Tit74] I. R. Titze. "The human vocal cords: A mathematical model, Part II," *Phonetica*, vol. 29, pp. 1–21, 1974.

[Tit94] I. R. Titze. *Principles of Voice Production*. Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[TS75] I. R. Titze and W.J. Strong. "Normal modes in vocal cord tissues," *J. Acous. Soc. Am.*, vol. 57, no. 3, pp. 736–744, 1975.

[Ult02] UltraVoice^{TM}. "http://www.ultravoice.com/how.htm," 2002.

[Väl95] V. Välimäki. *Discrete-time modeling of acoustic tubes using fractional delay filters*. Dissertation for the degree of doctor of technology, Laboratory of Acoustics and Audio Signal Processing, Faculty of Electrical Engineering, Helsinki University of Technology, Espoo, Finland, December 1995.

[WF89] J. Westbury and O. Fujimura. "An articulatory characterization of contrastive emphasis," *J. Acous. Soc. Am.*, vol. 85, no. Suppl. 1, pp. S98, 1989.

[WGKH98] D. H. Whalen, B. Gick, M. Kumada, and K. Honda. "Cricothyroid activity in high and low vowels: Exploring the automaticity of intrinsic F0," *Journal of Phonetics*, vol. 27, pp. 125–142, 1998.

[WGL99] D. H. Whalen, B. Gick, and P. S. LeSourd. "Intrinsic F0 in Passamaquoddy vowels," in *Papers from the 30th Algonquian Conference*, D. H. Pentland, Ed., University of Manitoba, Winnipeg, 1999, pp. 417–428.

[WL95] D. H. Whalen and A. G. Levitt. "The universality of intrinsic F0 of vowels," *Journal of Phonetics*, vol. 23, pp. 249–366, 1995.

[WMW84] R. L. Whitehead, D. E. Metz, and B. H. Whitehead. "Vibratory patterns of the vocal folds during pulse register phonation," *J. Acous. Soc. Am.*, vol. 75, no. 4, pp. 1293--1296, April 1984.

[Zee80] E. Zee. "Tone and vowel quality," *Journal of Phonetics*, vol. 8, pp. 247–258, 1980.

[Zem88] W.R. Zemlin. *Speech and Hearing Science: Anatomy & Physiology.* Prentice-Hall, Englewood Cliffs, 3rd edition, 1988.

[ZG89] P. A. Zawadzki and H. R. Gilbert. "Vowel fundamental frequency and articulator position," *J. Phonetics*, vol. 17, pp. 159–166, 1989.